

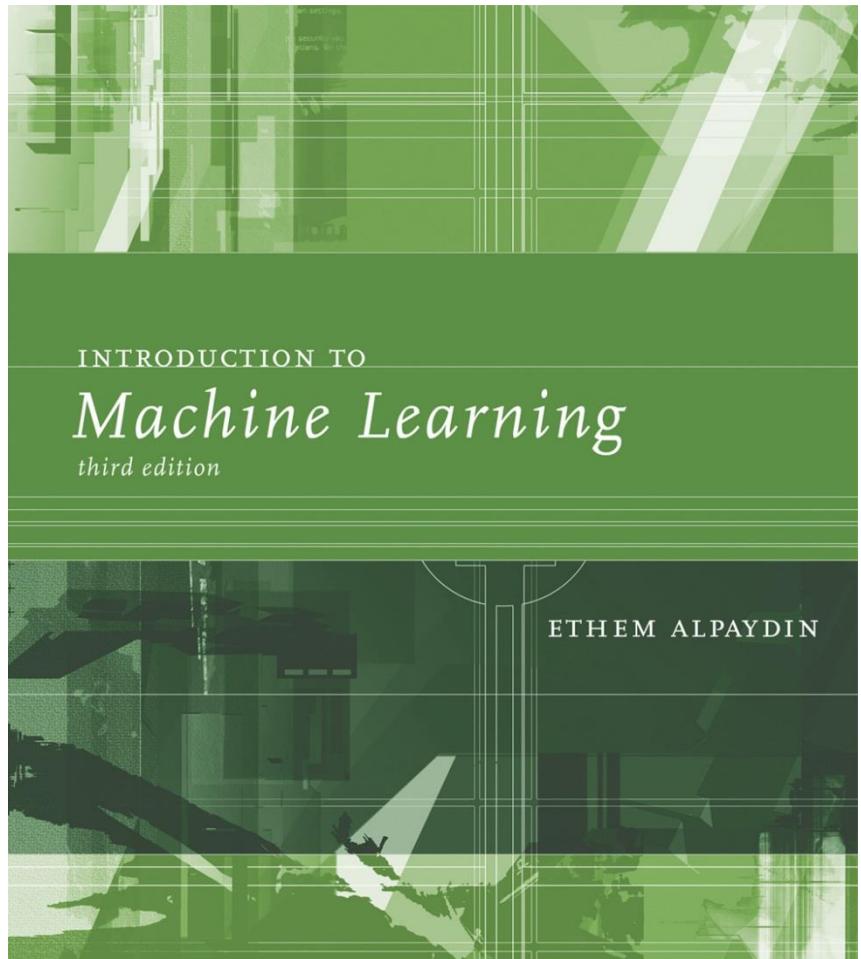
یادگیری ماشین: معرفی

سید ناصر رضوی n.razavi@tabrizu.ac.ir

۱۳۹۷

منابع و مراجع (۱)

۲



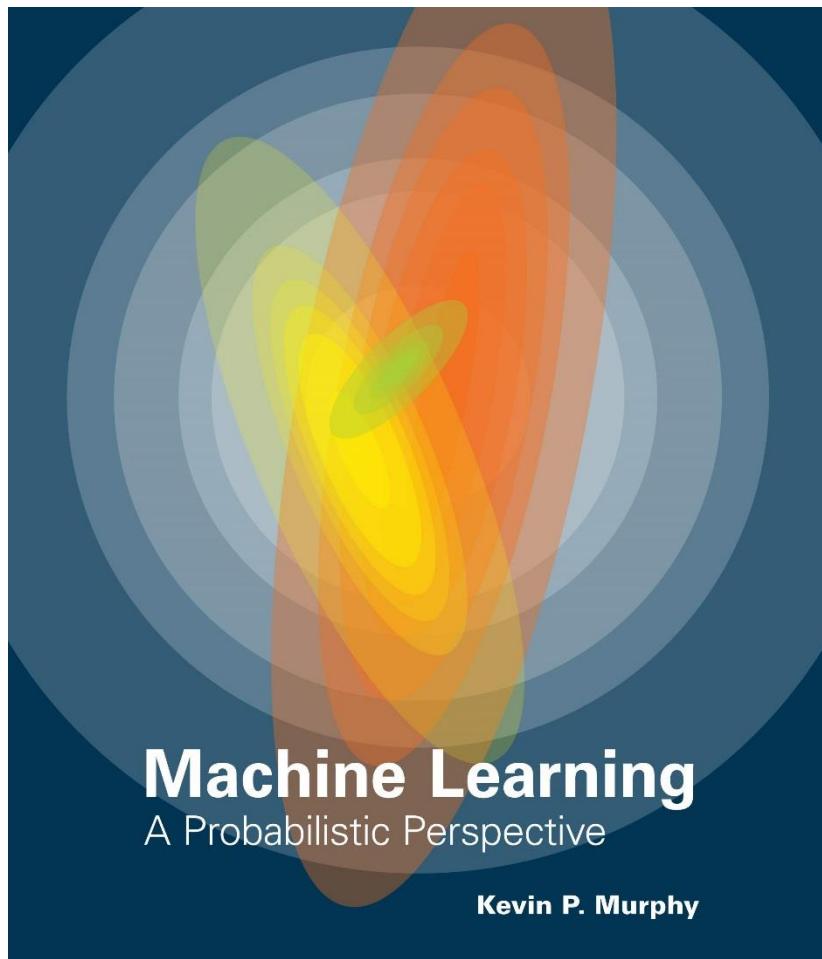
- مقدمه‌ای بر یادگیری ماشین.
[آلپایدین، ویراست سوم؛ ۲۰۱۴]

- یادگیری ماشین: یک دیدگاه احتمالاتی.
[کوین مورفی، ۲۰۱۲]

- شناسایی الگو و یادگیری ماشین.
[کریستوفر بیشاپ، ۲۰۰۶]

منابع و مراجع (۲)

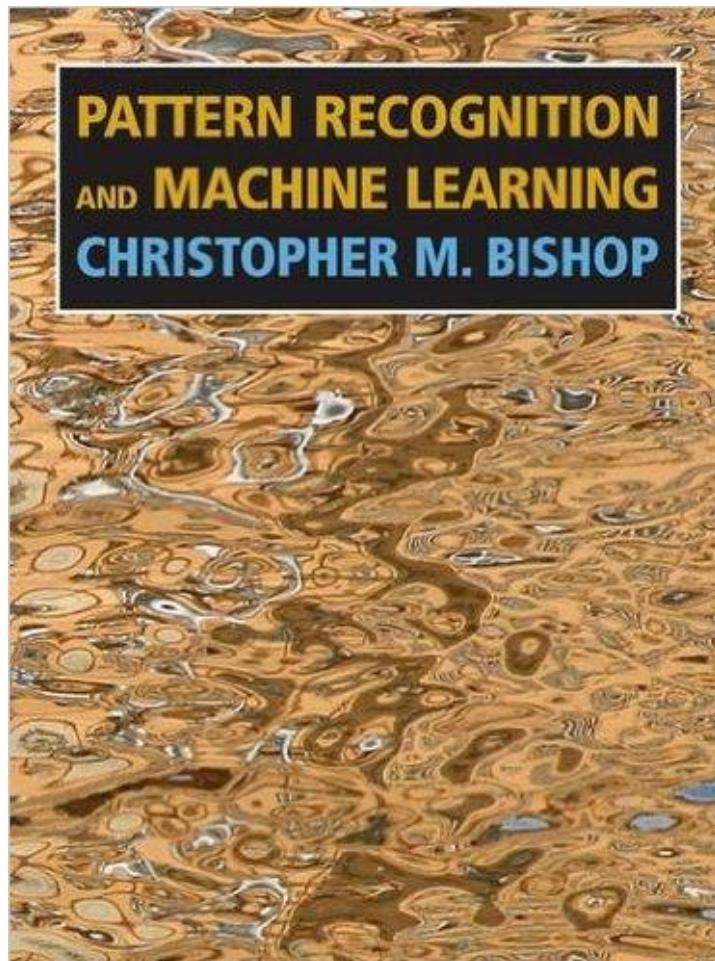
۳



- مقدمه‌ای بر یادگیری ماشین.
[آلپایدین، ویراست سوم؛ ۲۰۱۴]
- یادگیری ماشین: یک دیدگاه احتمالاتی.
[کوین مورفی، ۲۰۱۲]
- شناسایی الگو و یادگیری ماشین.
[کریستوفر بیشاپ، ۲۰۰۶]

منابع و مراجع (۳)

۴



- مقدمه‌ای بر یادگیری ماشین.
[آلپایدین، ویراست سوم؛ ۲۰۱۴]
- یادگیری ماشین: یک دیدگاه احتمالاتی.
[کوین مورفی، ۲۰۱۲]
- شناسایی الگو و یادگیری ماشین.
[بیشاپ، ۲۰۰۶]

پیش‌نیازها

۵

روش‌های تحلیل و طراحی الگوریتم‌ها

■ تحلیل پیچیدگی محاسباتی الگوریتم‌های یادگیری

جبر خطی

■ ماتریس‌ها، بردارها، عملیات ماتریسی و دستگاه معادلات خطی

■ ماتریس وارون، بردارهای ویژه، مرتبه ماتریس، تجزیه مقادیر منفرد

حساب چند متغیره

■ مشتق، انتگرال، صفحات مماس

احتمالات

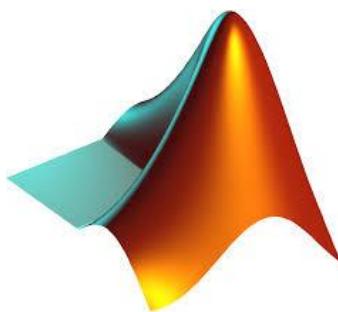
■ متغیرهای تصادفی، مقدار مورد انتظار، واریانس و ...

ارزیابی

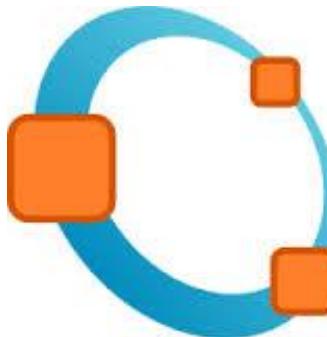
۶



- تمرین‌ها [۴۰٪]
- مباحث نظری
- برنامه‌نویسی
- امتحان پایان‌ترم [۵۰٪]



MATLAB



OCTAVE

فهرست مطالب

۷

- یادگیری ناظارت شده.
- رگرسیون - رگرسیون خطی تک متغیره و چند متغیره
- دسته‌بندی - رگرسیون لجستیک، شبکه‌های عصبی، ماشین‌های بردار پشتیبان
- یادگیری بدون ناظارت.
- خوشبندی
- یادگیری تقویتی.
- برنامه‌نویسی با استفاده از زبان پایتون (یا اکتاو).
- توصیه‌های عملی در استفاده از الگوریتم‌های یادگیری ماشین.

چند نقل قول

۸

«هر گام رو به چلو در جهت یادگیری ماشین دهها برابر مایکروسافت ارزش دارد»



بیل گیتس - مدیر مایکروسافت

«نسل بعدی اینترنت پیزی به هز یادگیری ماشین نیست»



تونی تدر - مدیر اسبق دارپا

«یادگیری ماشین در نهایت به یک انقلاب واقعی منجر خواهد شد»



گرگ پاپادopoulos - مدیر اسبق سان

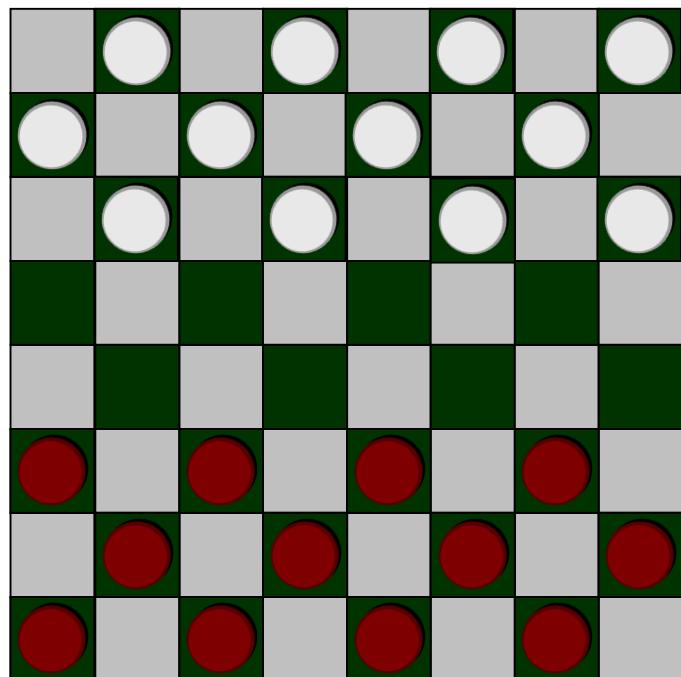
یادگیری ماشین پیشست؟

یادگیری ماشین: تعاریف

۱۰

□ آرتو ساموئل. [۱۹۵۹]

« یک حوزه مطالعاتی که به ماشین‌ها توانایی یادگیری می‌دهد، بدون این که نیاز باشد این ماشین‌ها به طور صریح برنامه‌نویسی شوند. »



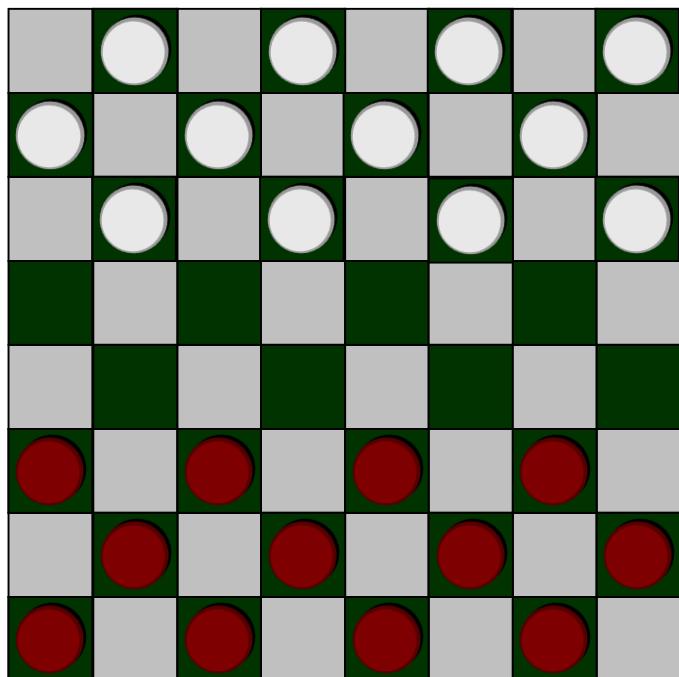
□ بازی چکرز. [ساموئل، دهه ۵۰]

یادگیری ماشین: تعاریف

۱۱

□ تام میشل. [۱۹۹۸]

« با داشتن یک وظیفه مانند T و یک معیار کارایی مانند P ، می‌گوییم یک برنامه کامپیوترا از تجربه‌ی E یاد می‌گیرد اگر معیار کارایی آن برنامه یعنی P برای انجام وظیفه T با استفاده از تجربه E بهبود یابد. »



□ مثال. بازی چکرز

□ **وظیفه:** انجام بازی چکرز

□ **تجربه:** هزاران هزار بار بازی در مقابل خود

□ **معیار کارایی:** تعداد دفعات برد در برابر رقبای جدید

مثال: تشفیض هرزنامه

۱۲

The screenshot shows a Gmail inbox with 7 messages in the inbox. A specific message is highlighted, which is a Google Account recovery email. The message subject is "Google Account recovery phone number changed". The message body states that the recovery phone number for the account has been changed and provides a link for account recovery. The message is from "accounts-noreply@google.com" and was sent at 4:43 PM (0 minutes ago). The message is marked as spam.

New! Gmail's mobile apps just got updated on [Google Play](#) and the [Apple App Store](#). [Dismiss](#)

Report spam

(3 people unfriended you) - [www.UnfriendApp.com](#) - Free FB tool that shows you who unfriended you!

Why this ad?

Ads – Why these ads?

Email Auto-Responders
Quickly Engage New Leads & Contacts Email Auto-Responder.
Free Account!
[www.Contactology.com/AutoRespon](#)

Do You Carry Concealed?
Know Your Rights & Get Your Free Concealed Carry Report Today!
[USConcealedCarry.net](#)

Try GFI® VIPRE Free
Award-winning antivirus software.
Download a free 30-day trial now.
[www.VIPREbusiness.com](#)

2013 Grants
Grant Funding May Be Available
See If You Qualify!
[www.ClassesUSA.com](#)

Note: This email address cannot accept replies.

Sincerely,
The Google Accounts Team

© 2012 Google Inc. 1600 Amphitheatre Parkway, Mountain View, CA 94043

You have received this mandatory email service announcement to update you about important changes to your Google product or account.

مثال: تشخیص هرزنامه

۱۳

□ مثال. تشخیص هرزنامه

فرض کنید برنامه ایمیل شما به شما امکان می‌دهد که ایمیل‌های دریافتی خود را به عنوان هرزنامه علامت بزنید و بر این اساس یاد می‌گیرد که چگونه هرزنامه‌ها را بهتر فیلتر کند.

□ **وظیفه:** دسته‌بندی ایمیل‌ها به عنوان هرزنامه یا ایمیل.

□ **تجربه:** نظارت بر این که شما کدام ایمیل‌ها را به عنوان هرزنامه علامت می‌زنید.

□ **معیار کارایی:** تعداد ایمیل‌هایی که به درستی دسته‌بندی شده‌اند.

انواع روش‌های یادگیری ماشین

۱۴

- یادگیری ماشین. بهبود عملکرد ماشین در انجام یک وظیفه با کسب تجربه.
- س. یک ماشین از کجا می‌تواند بفهمد عملکردنش بهبود یافته است؟
- می‌توانیم به ماشین پاسخ درست را برای چند نمونه محدود از ورودی‌ها بدھیم به این امید که بتواند آن را برای نمونه‌های دیگر تعمیم دهد -- **یادگیری ناظارت شده**
- می‌توانیم به ماشین بگوییم پاسخش تا چه میزان درست بوده (مثلاً با دادن یک امتیاز) و خود ماشین مسئول یافتن پاسخ‌های درست است -- **یادگیری تقویتی**
- ممکن است هیچ اطلاعاتی در مورد پاسخ درست به ماشین ندهیم و تنها از ماشین بخواهیم ورودی‌هایی را که دارای وجه مشترک هستند پیدا کند -- **یادگیری بدون ناظارت**

يادگيري نظارت شده

یادگیری نظارت شده

۱۶

□ ورودی. یک مجموعه آموزشی که در آن به ازای هر ورودی پاسخ درست داده شده است.

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$

مجموعه آموزشی



□ هدف. یافتن یک تقریب مناسب برای نگاشت زیر:

$$f: X \rightarrow Y$$

□ مثال.

□ تشخیص هرزname: نگاشت ایمیل‌ها به مجموعه {هرزنامه، غیرهرزنامه}

□ تشخیص ارقام: نگاشت یک مجموعه از پیکسل‌ها به مجموعه {۰، ۱، ۲، ...، ۹}

□ تشخیص سرطان: نگاشت داده‌های پزشکی به مجموعه {بدخیم، خوشخیم}

مثال: تشفیض هرزنامه

۱۷

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret. ...



To be removed from future mailings, simply reply to this message and put "remove" in the subject.

99 million email addresses for only \$99



Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.



ورودی. ایمیل

خروجی. هرزنامه، غیرهرزنامه

مثال: تشفیض ارقام دستنویس

۱۸

۸	۳	۹	۳	۸	۵	۸	۵	۶	۵
۹	۴	۹	۵	۷	۱	۷	۶	۱	۱
۶	۸	۳	۶	۸	۸	۸	۱	۱	۴
۴	۹	۵	۰	۱	۲	۱	۴	۵	۳
۷	۲	۷	۷	۶	۳	۱	۱	۲	۱
۳	۲	۷	۰	۴	۶	۰	۸	۱	۸
۶	۰	۷	۴	۱	۱	۷	۴	۲	۱
۲	۹	۵	۳	۷	۴	۱	۰	۵	۸
۳	۵	۵	۷	۶	۵	۹	۹	۹	۳
۱	۹	۹	۶	۱	۲	۱	۳	۶	۷

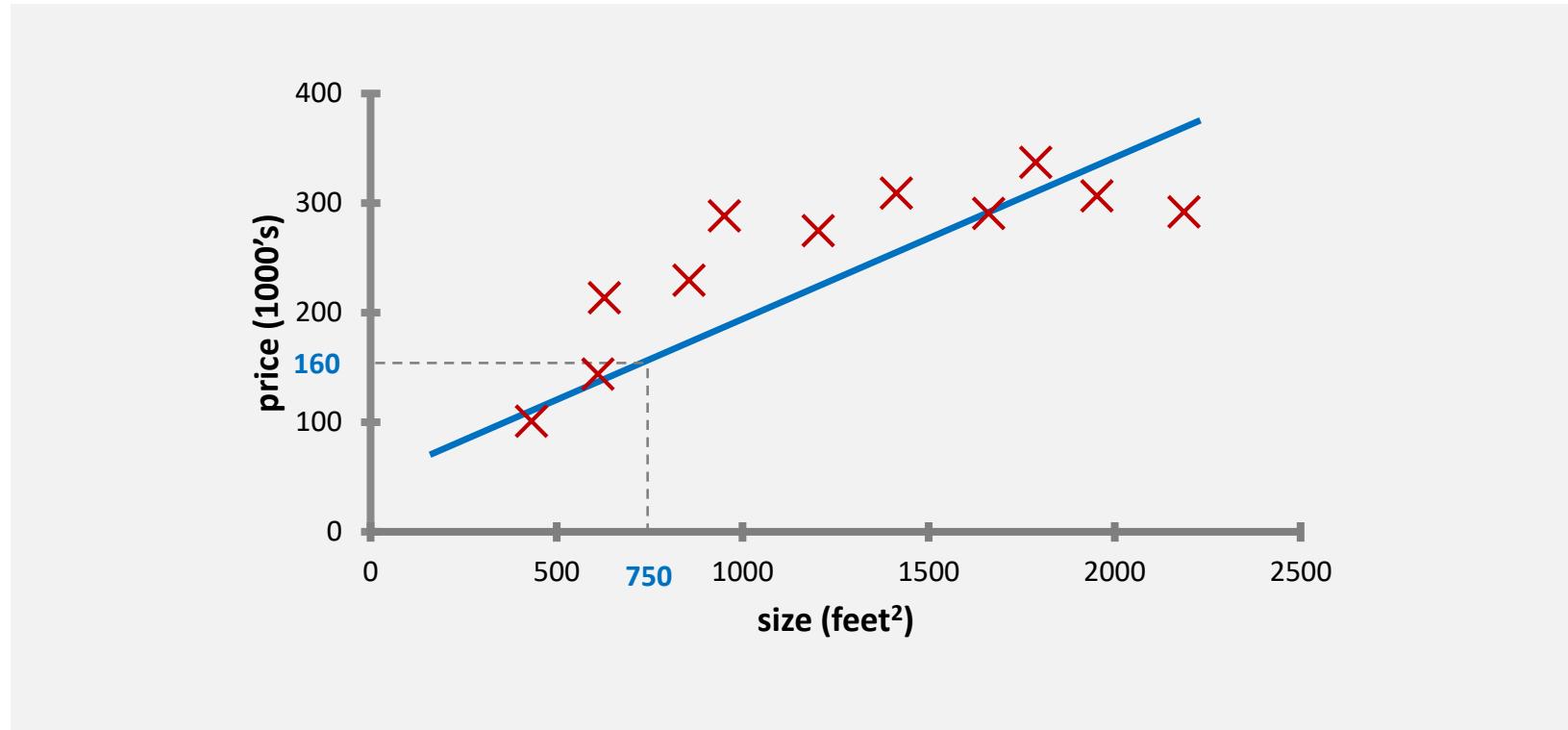
- ورودی. تصویر یک رقم
- خروجی. یک رقم

مثال: قیمت‌گذاری یک خانه

۱۹

ورودی. اندازه خانه [بر حسب فوت مربع]

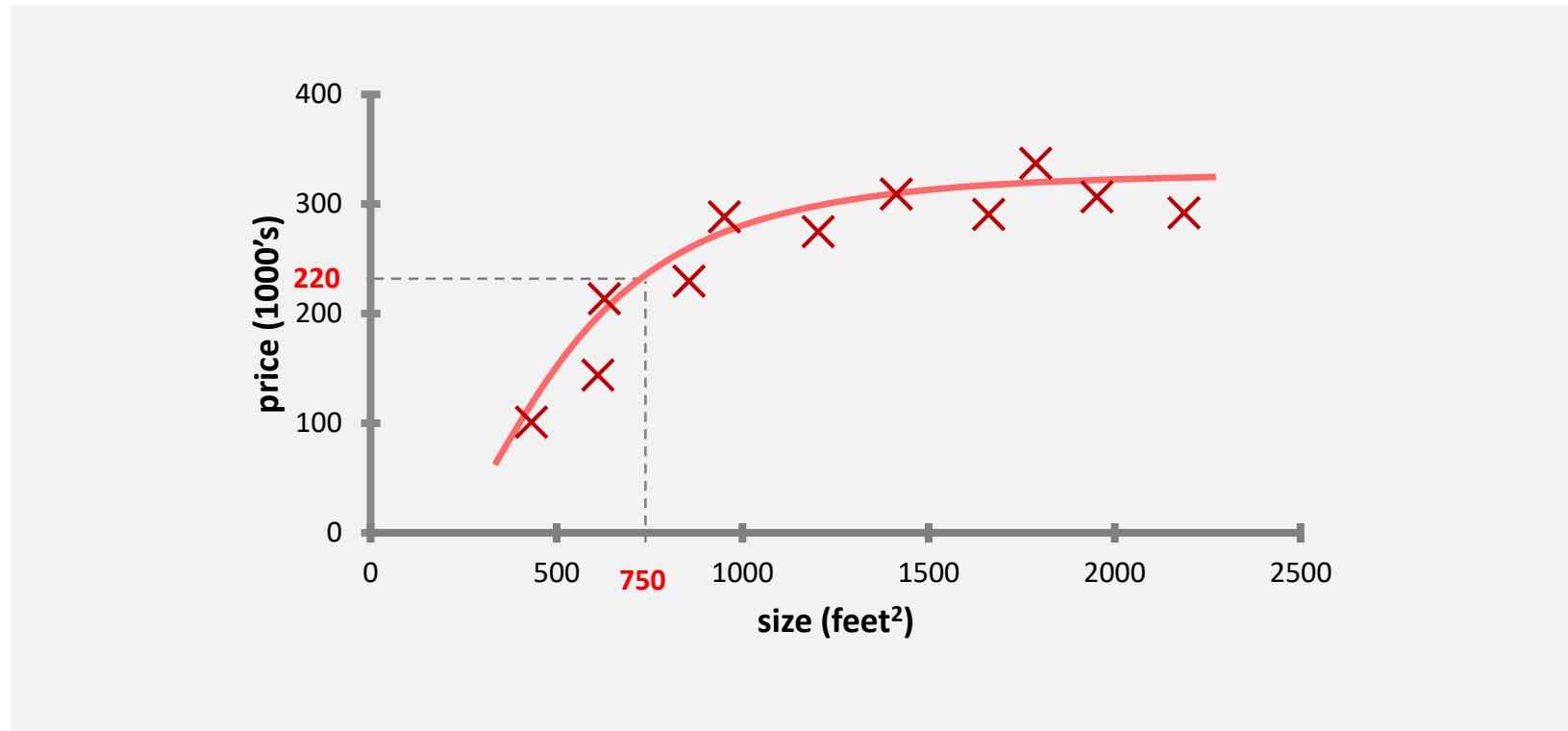
خروجی. قیمت تخمینی



مثال: قیمت‌گذاری یک خانه

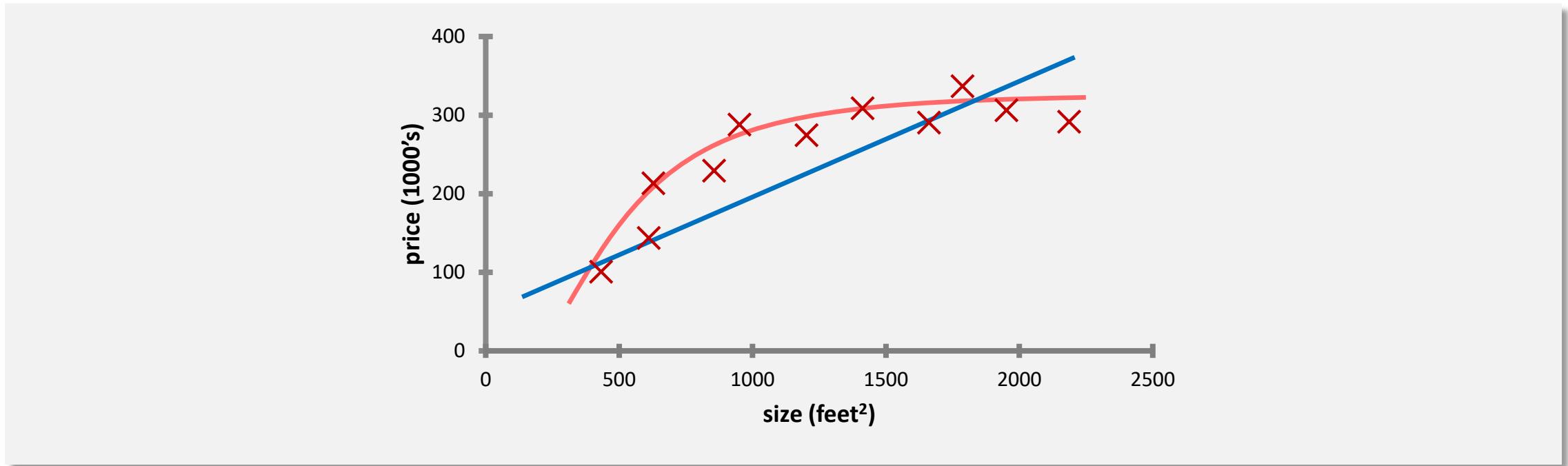
۲۰

س. کدام یک بهتر است؟ یک تابع خطی یا یک تابع درجه دوم؟



مثال: قیمت‌گذاری یک خانه

۲۱



□ رگرسیون.

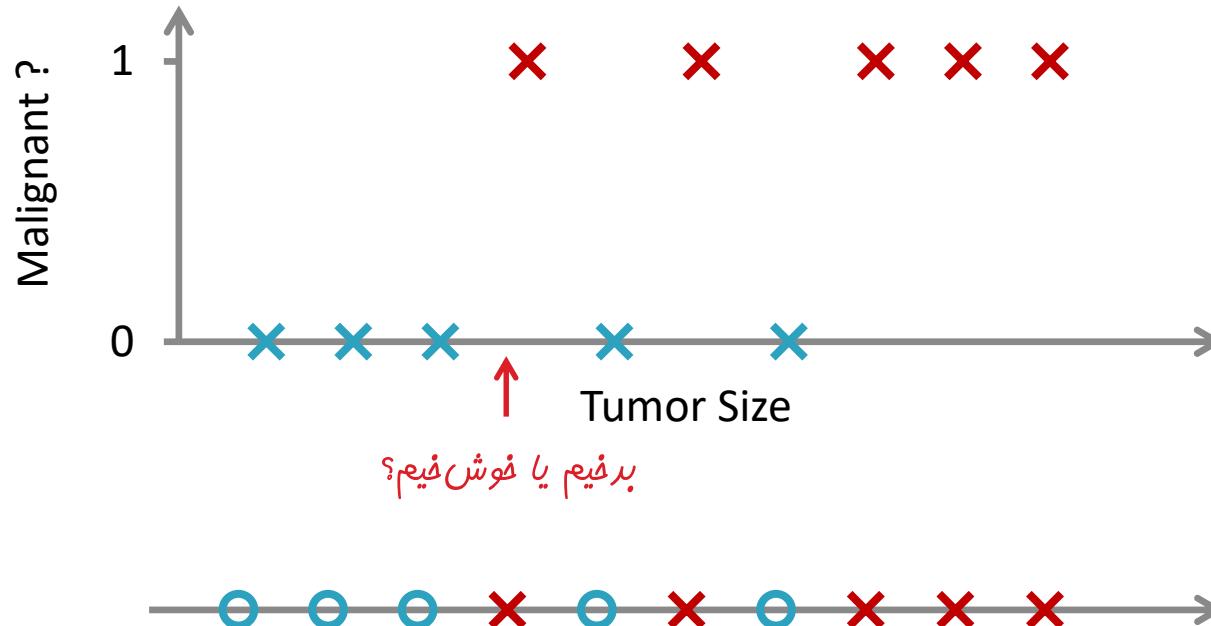
پیش‌بینی کمیت‌هایی با مقادیر پیوسته (مانند قیمت یک خانه)

□ یادگیری نظارت شده.

به ازای هر نمونه آموزشی، «پاسخ درست» داده شده است.

مثال: تشفیم نوع سرطان (بدفیم، خوش‌فیم)

۲۲

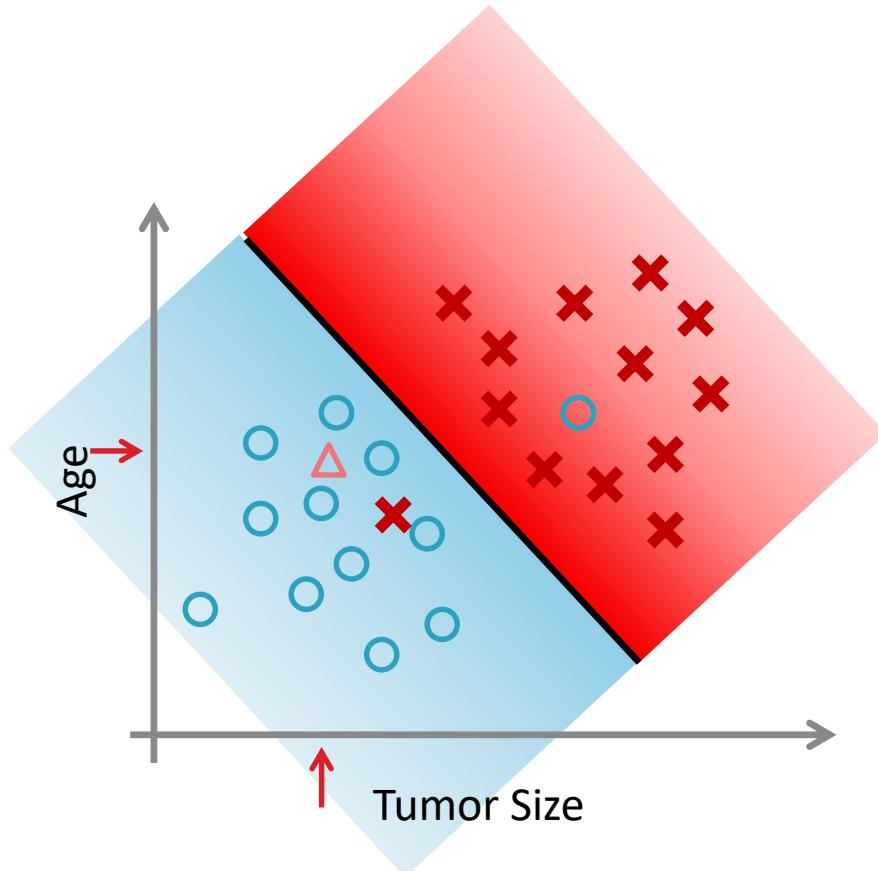


□ کلاس‌بندی.
پیش‌بینی کمیت‌هایی با مقادیر گسته (مانند صفر و یک).

□ یادگیری نظارت شده.
به ازای هر نمونه آموزشی، «پاسخ درست» داده شده است.

مثال: تئشیم نوع سرطان (بدخیم، خوشخیم)

۲۳



ویژگی‌های دیگر.

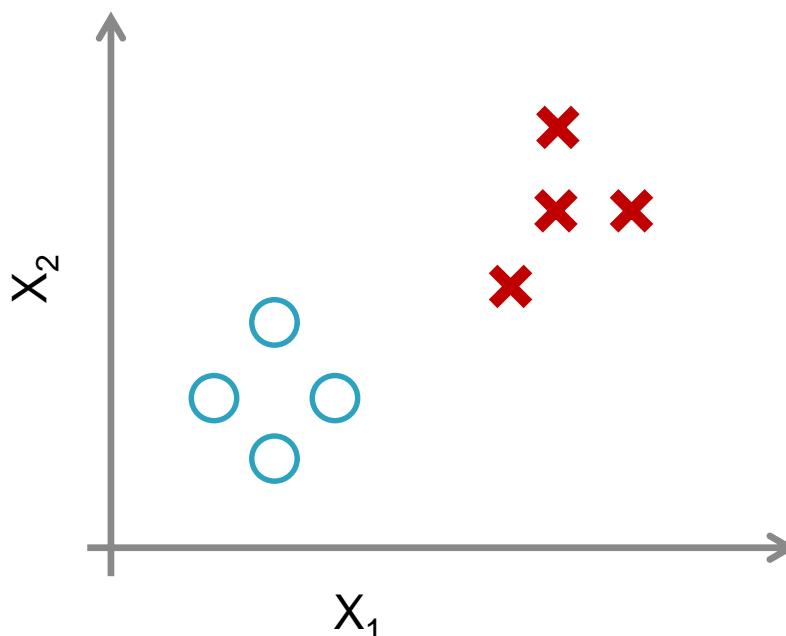
- یکنواختی اندازه سلول‌ها
- یکنواختی شکل سلول‌ها
- ... ○

يادگيري بدون نظارت

یادگیری نظارت شده

۲۵

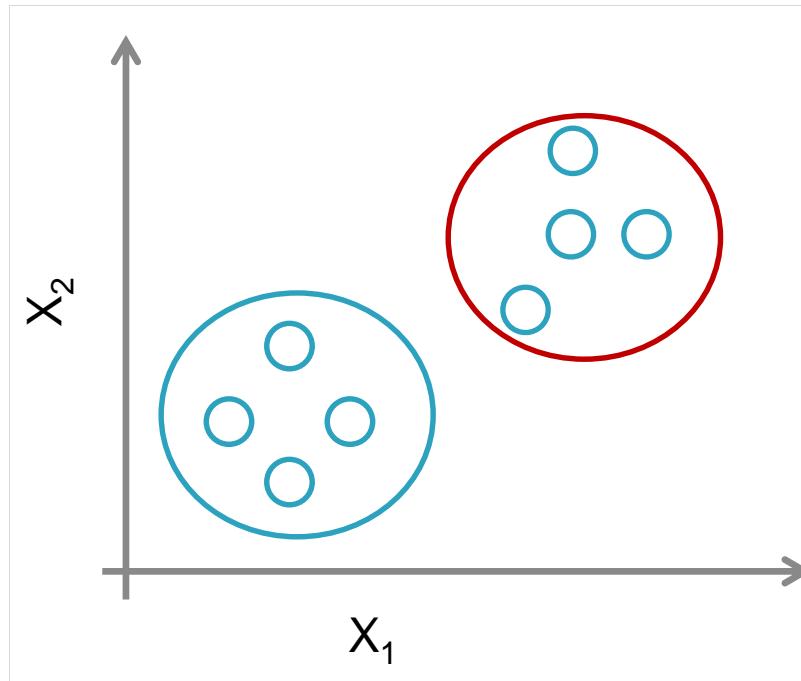
یادگیری نظارت شده. به ازای هر نمونه، پاسخ درست داده شده است. □



یادگیری بدون نظارت

۲۶

□ یادگیری بدون نظارت. هیچ گونه اطلاعاتی در مورد پاسخ‌های درست داده نشده است!



□ هدف. تشخیص ساختار در داده‌های ورودی (گروه‌بندی داده‌های مشابه).

کاربرد خوشنودی: گروه‌بندی اخبار مرتب

۲۷

The screenshot shows the Google News interface. On the left, there's a sidebar with categories: Top Stories, Starred, San Francisco Bay Area, World, U.S., Business, Sci/Tech, More Top Stories, Spotlight, Health, Sports, and Entertainment. Below that are links for All news, Headlines, and Images.

The main content area has several sections:

- Top Stories:** A large box containing an article about the White House denying a Tea Party-focused ad campaign, followed by other news items like US stocks climbing and BP oil well issues.
- Recent:** A list of recent news items from various sources, such as the recession ending, Hurricane Igor, and a San Francisco Bay Area story about Clorox.
- San Francisco Bay Area - Edit:** A section specifically for local news in the Bay Area.

A red box highlights the "BP Oil Well, Site of National Catastrophe, Dies at One" article under the Top Stories section. This article discusses the Deepwater Horizon explosion and its aftermath.

کاربرد خوشنودی افبار مرتبط

۲۸

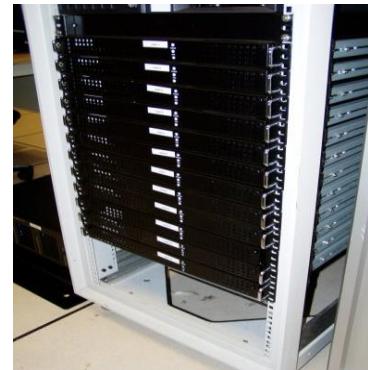
The diagram illustrates the interconnectedness of news stories about the Deepwater Horizon oil spill. It shows four screenshots:

- Google News:** Shows a search result for "BP Oil Well, Site of National Catastrophe, Dies at One". A red box highlights this story.
- The Wall Street Journal - THE SOURCE:** Shows a story titled "BP Kills Macondo, But Its Legacy Lives On".
- CNN:** Shows a story titled "Allen: Well is dead, but much Gulf Coast work remains".
- The Guardian:** Shows a story titled "BP oil spill cost hits nearly \$10bn".

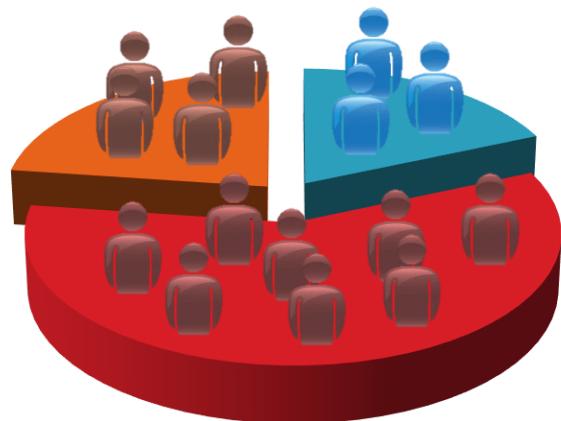
Three red arrows point from the highlighted story in Google News to the equivalent stories in The Wall Street Journal, CNN, and The Guardian, demonstrating how a single news item can trigger multiple coverage across different media outlets.

چند کاربرد دیگر از یادگیری بدون نظارت

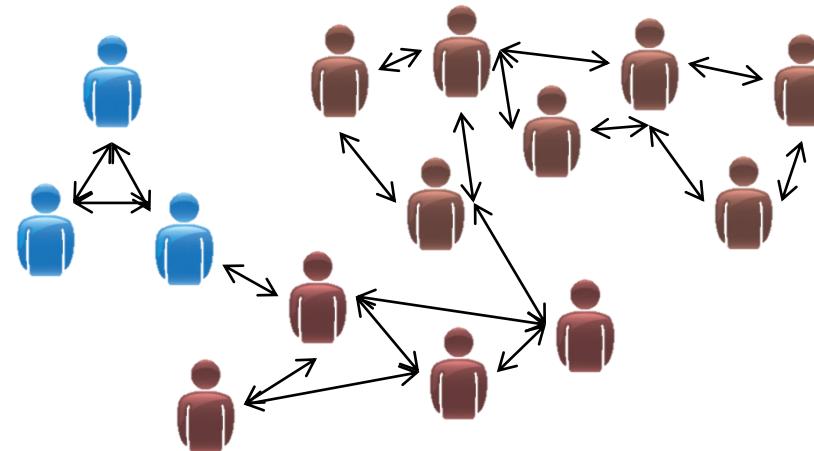
۲۹



سازماندهی کلاسترها مهاسباتی (مرکز داده‌ها)



بفشنبدی بازار



تحلیل شبکه‌های اجتماعی

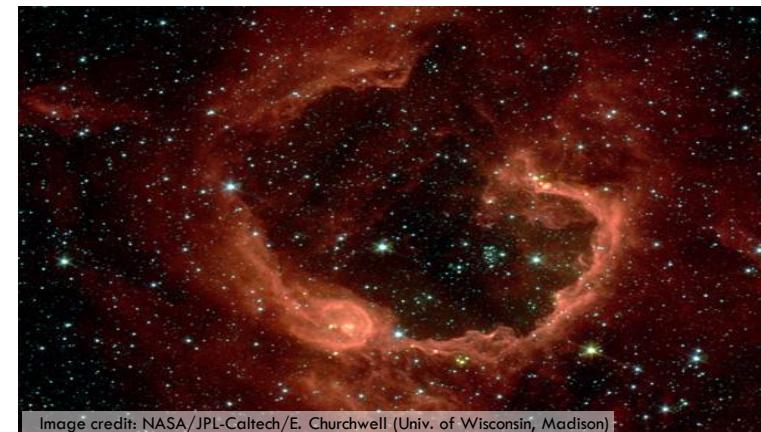
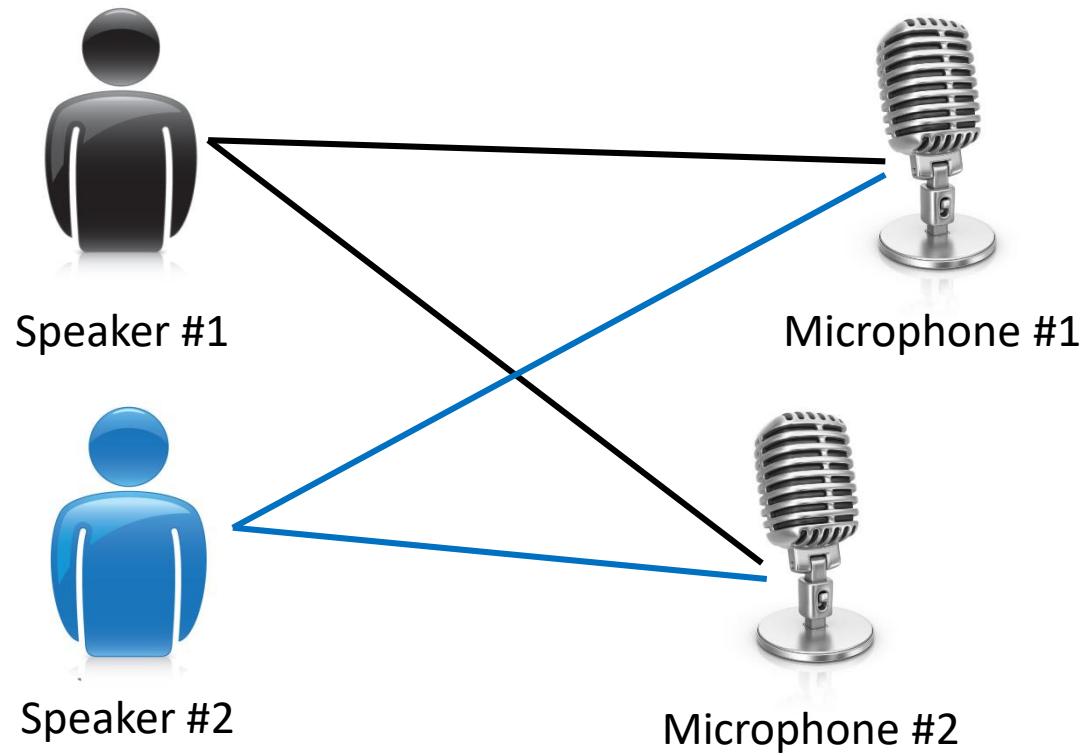


Image credit: NASA/JPL-Caltech/E. Churchwell (Univ. of Wisconsin, Madison)

تحلیل داده‌های ستاره‌شناسی (نحوه تشکیل کوکشان‌ها)

مسئله جشن کوکتل

۳۰



مسئله جشن کوکتل

۳۱

Microphone #1: 

Output #1: 

Microphone #2: 

Output #2: 

Microphone #1: 

Output #1: 

Microphone #2: 

Output #2: 

الگوريتم مسئله جشن کوکتل

۳۲

□ کد اکتاو.

```
[W, s, v] = svd((repmat(sum(x .* x, 1), size(x, 1), 1) .* x) * x');
```

پرسش کلاسی

۳۳

□ برای کدام یک از مسائل داده شده زیر باید از یک الگوریتم یادگیری بدون نظارت استفاده شود؟

- توسعه یک برنامه برای فیلتر کردن هرزنامه‌ها با داشتن تعدادی ایمیل معمولی و تعدادی هرزنامه گروه‌بندی یک مجموعه از مقالات جدید یافته شده در وب بر اساس موضوع
- گروه‌بندی مجموعه‌ای از مشتری‌ها در چند بخش مختلف بازار با داشتن یک پایگاه داده در مورد مشتریان
- تشخیص دیابت در بیماران جدید با داشتن داده‌های مربوط به تعدادی فرد سالم و دیابتی

یادگیری نظرات شده: (کرسیون

سید ناصر رضوی n.razavi@tabrizu.ac.ir

۱۳۹۷

فهرست مطالب

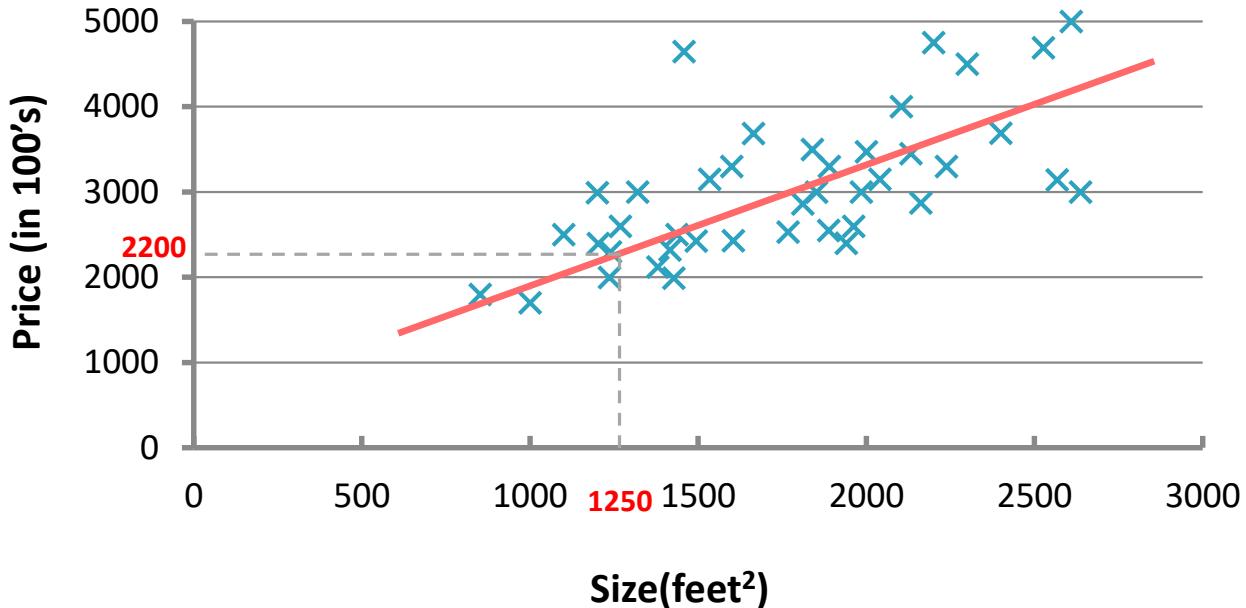
۲

- رگرسیون.
- رگرسیون خطی تک متغیره و چند متغیره
- گرادیان کاہشی.
- معادله نرمال.
- رگرسیون با وزن دهی محلی.
- تفسیر احتمالاتی رگرسیون.
- تخمین بیشترین درست نمایی.

ڪرسيون ڌطي تک مٿغیره

قیمت‌گذاری خانه

۴



رگرسیون.

پیش‌بینی کمیت‌هایی با مقادیر پیوسته. (مانند
قیمت یک خانه)

یادگیری نظارت شده.

به ازای هر نمونه آموزشی، «پاسخ درست» داده شده است.

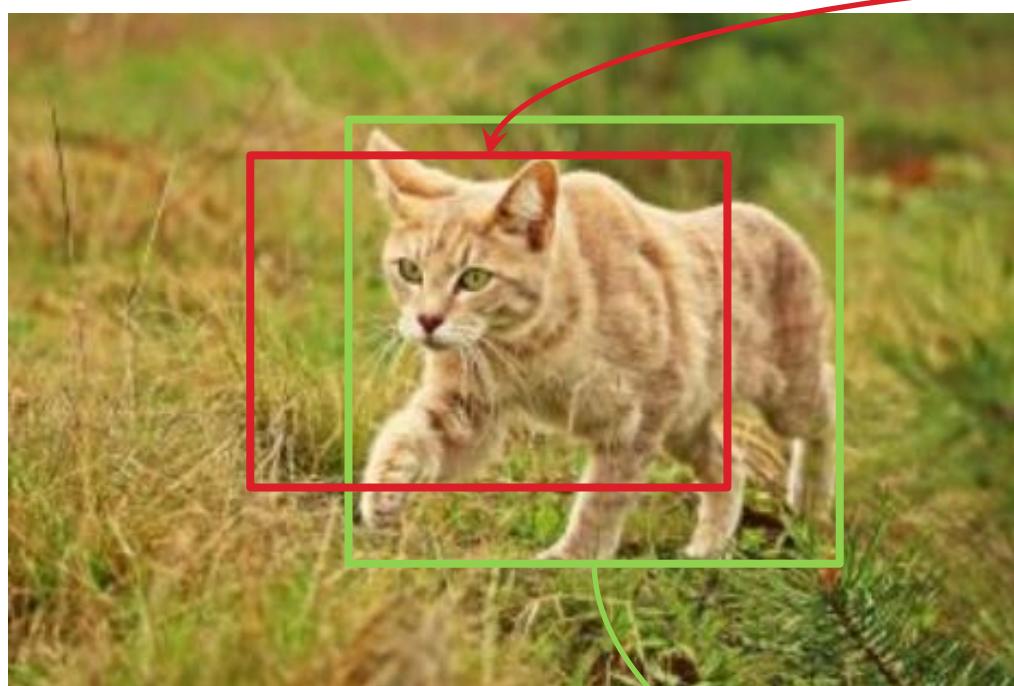
اگرسيون: شناسايي اشيا

۵

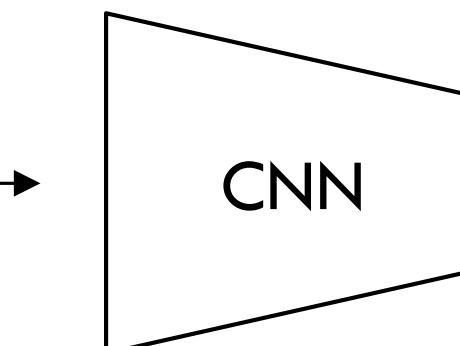


مکان‌یابی به عنوان (گرسيون

۶



پيش‌بيني



64.1
25.6
245.2
184.5

مجموع
مربعات
خطا

80.0
20.0
240.0
240.0

مقادير درست

نمادگذاری

۷

مجموعه آموزشی	متراز (فوت مربع) (x)	قیمت (در ۱۰۰۰ دلار) (y)
	۲۱۰۴	۴۶۰
	۱۴۱۶	۲۳۲
	۱۵۳۴	۳۱۵
	۸۵۲	۱۷۸
...

$$m = 47$$

(x, y) : یک نمونه آموزشی

$(x^{(i)}, y^{(i)})$: نمونه آموزشی i ام

m = تعداد نمونه‌های آموزشی

x = متغیر «ورودی»، ویژگی‌ها

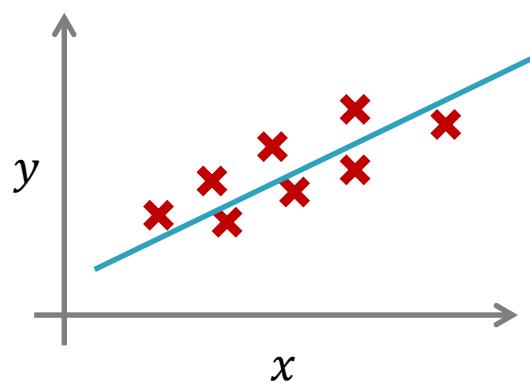
y = متغیر «خروجی»، متغیر «هدف»

بازنمایی مدل

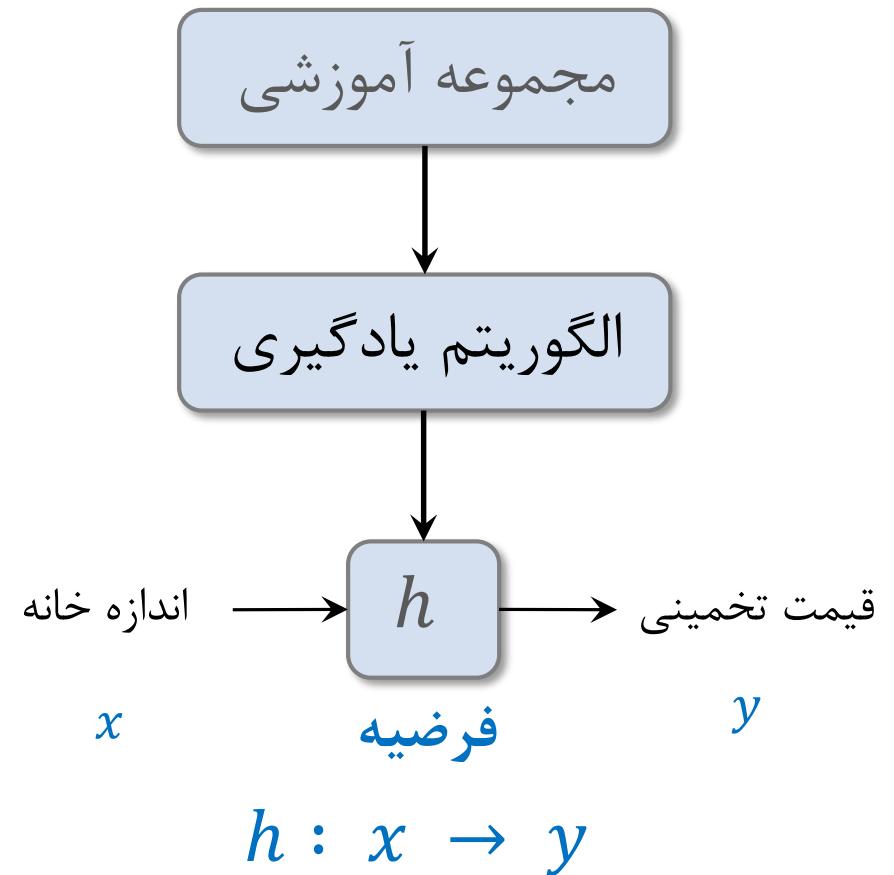
۸

نمایش فرضیه h

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$



رگرسیون خطی تک متغیره



تابع هزینه

ارزیابی فرضیه

۱۰

متراز (فوت مربع)	قیمت (در ۱۰۰۰ دلار)	مجموعه آموزشی
۴۶۰	۲۱۰۴	
۲۳۲	۱۴۱۶	
۳۱۵	۱۵۳۴	
۱۷۸	۸۵۲	
...	...	

$m = 47$

$h_{\theta}(x) = \theta_0 + \theta_1 x$

(θ_0, θ_1)

فرضیه:

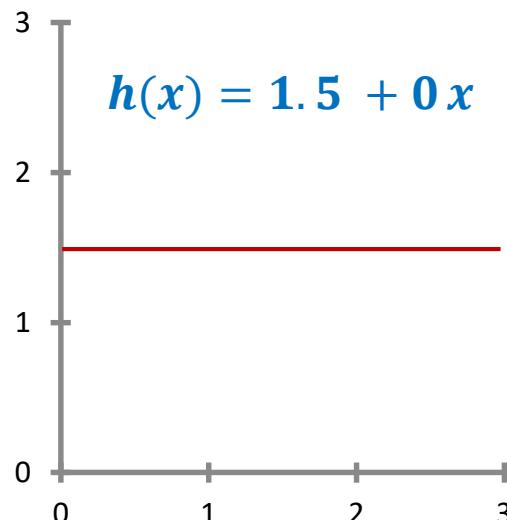
پارامترها:

س. مقدار پارامترها را چگونه باید انتخاب نمود؟

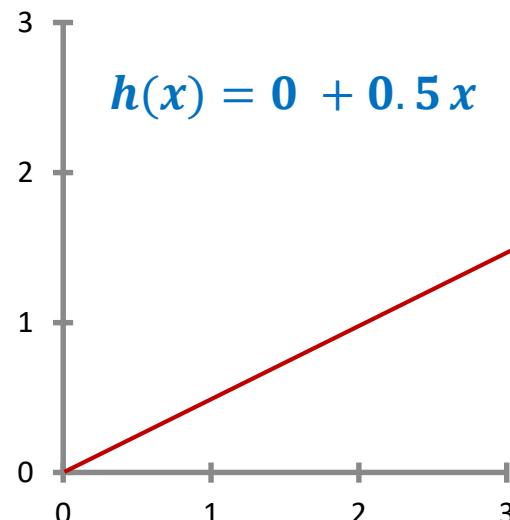
اڑیابی فرضیہ

۱۱

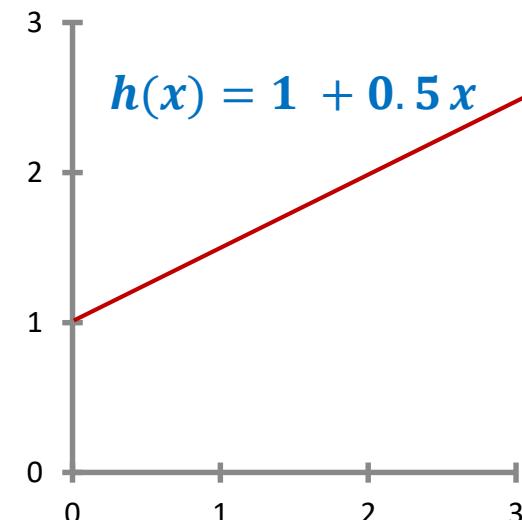
$$h_{\theta}(x) = \theta_0 + \theta_1 x$$



$$\begin{aligned}\theta_0 &= 1.5 \\ \theta_1 &= 0\end{aligned}$$



$$\begin{aligned}\theta_0 &= 0 \\ \theta_1 &= 0.5\end{aligned}$$



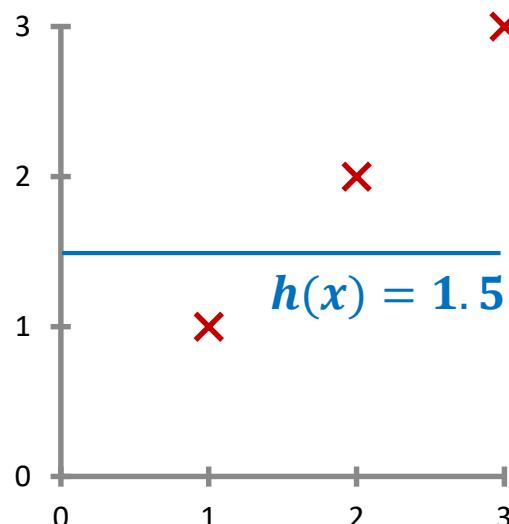
$$\begin{aligned}\theta_0 &= 1 \\ \theta_1 &= 0.5\end{aligned}$$

ارزیابی فرضیه

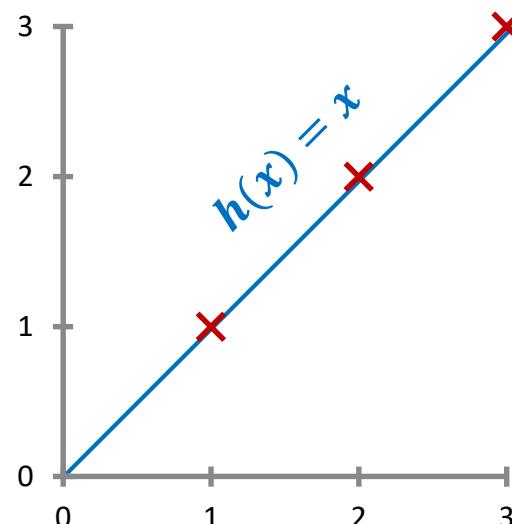
۱۲

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

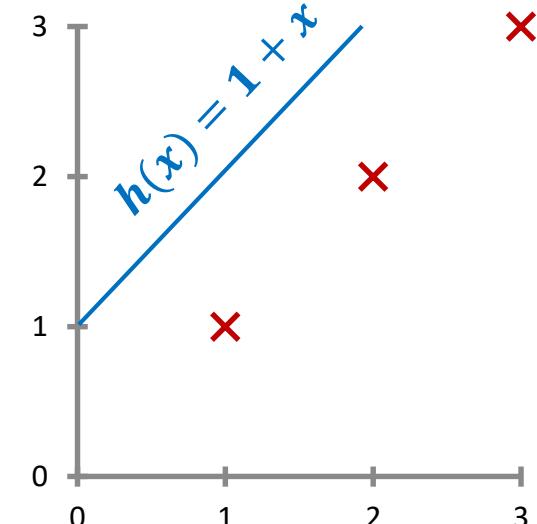
س. کدام فرضیه بهتر است؟



$$\begin{aligned}\theta_0 &= 1.5 \\ \theta_1 &= 0\end{aligned}$$



$$\begin{aligned}\theta_0 &= 0 \\ \theta_1 &= 1\end{aligned}$$

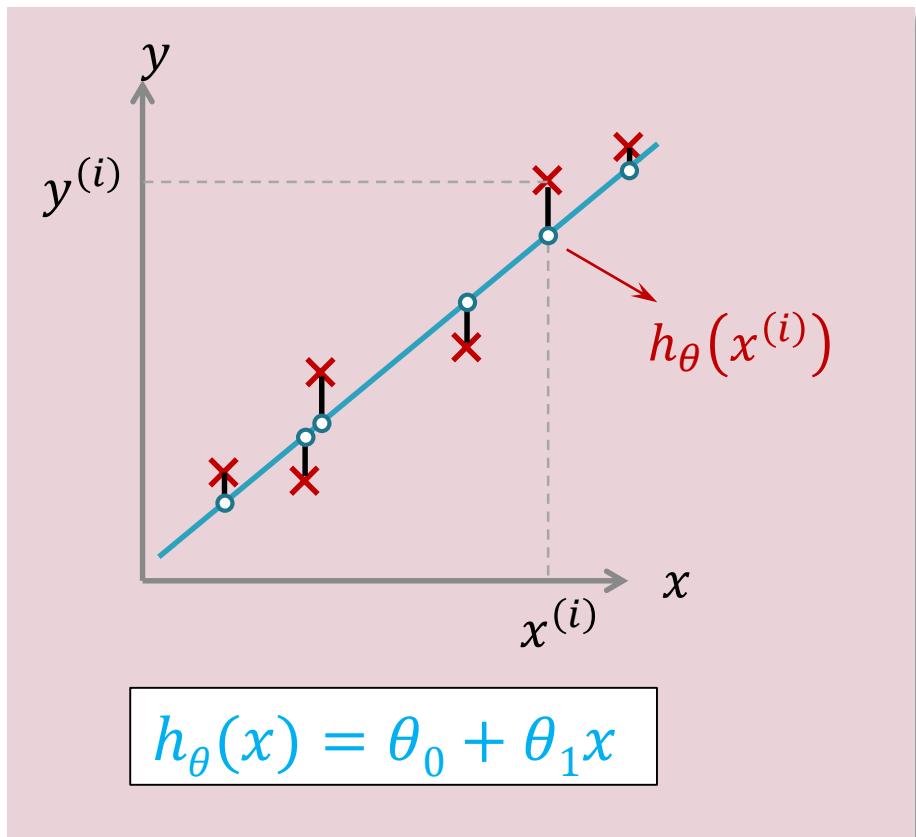


$$\begin{aligned}\theta_0 &= 1 \\ \theta_1 &= 1\end{aligned}$$

تابع هزینه

۱۳

□ ایده. انتخاب پارامترها به گونه‌ای که به ازای هر نمونه آموزشی مانند (x, y) ، مقدار $h_\theta(x)$ تا حد ممکن به مقدار y نزدیک باشد.



□ تابع هزینه. مجموع مربعات خطأ.

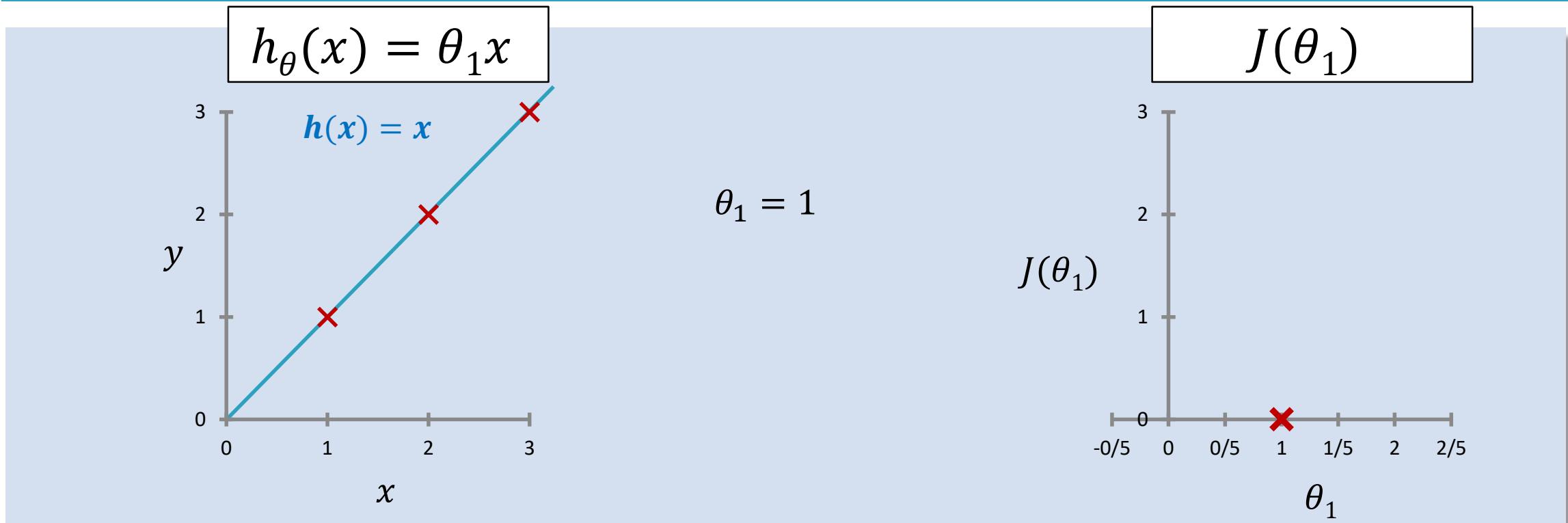
$$J(\theta_0, \theta_1) = \frac{1}{2} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

□ هدف.

$$\underset{\theta_0, \theta_1}{\text{minimize}} \ J(\theta_0, \theta_1)$$

تابع هزینه ساده شده

۱۴

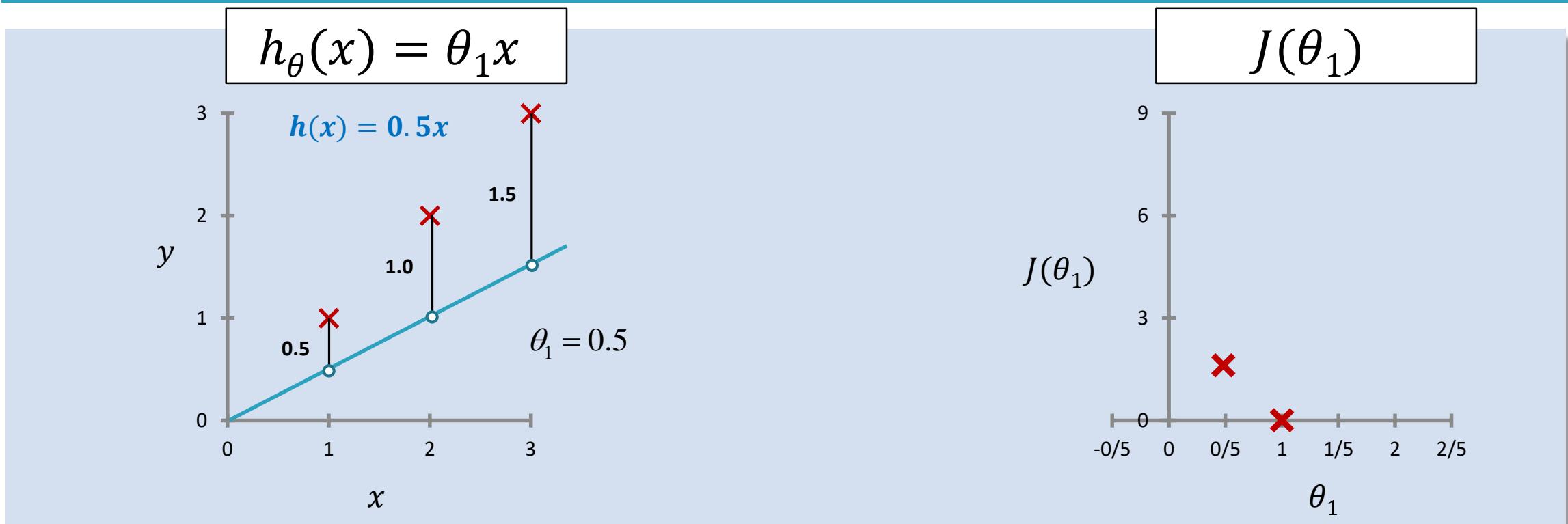


$$\begin{aligned}
 J(\theta_0, \theta_1) &= \frac{1}{2} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2 \\
 &= \frac{1}{2} \sum_{i=1}^m (x^{(i)} - y^{(i)})^2 = \frac{1}{2}(0^2 + 0^2 + 0^2) = 0
 \end{aligned}$$

$$J(1) = 0$$

تابع هزینه

١٨



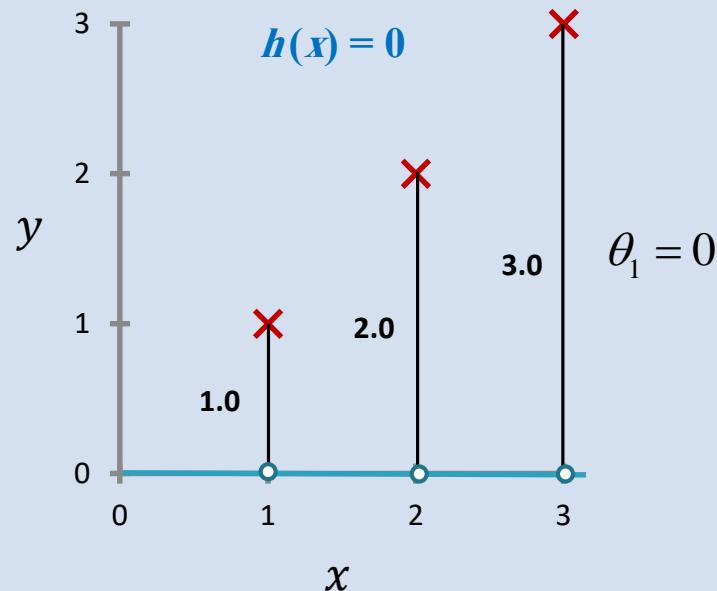
$$J(0.5) = \frac{1}{2}(0.5^2 + 1.0^2 + 1.5^2) = \frac{1}{2}(3.5) = 1.75$$

$$J(0.5) = 1.75$$

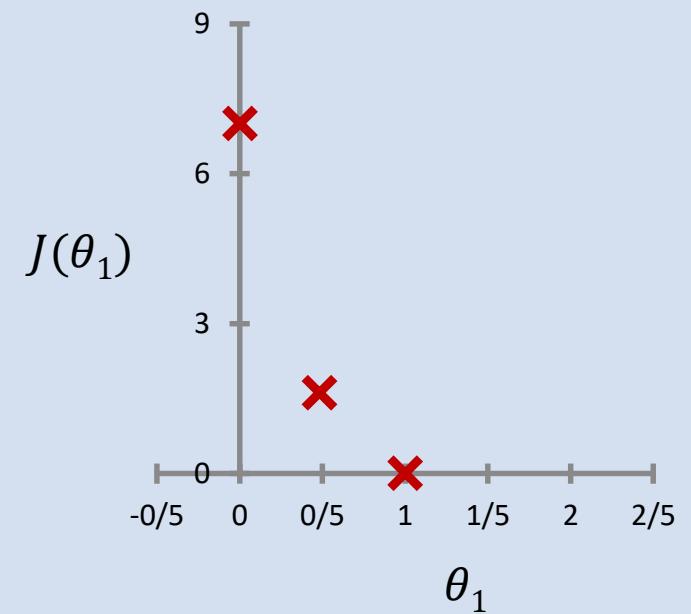
تابع هزینه

١٦

$$h_{\theta}(x) = \theta_1 x$$



$$J(\theta_1)$$



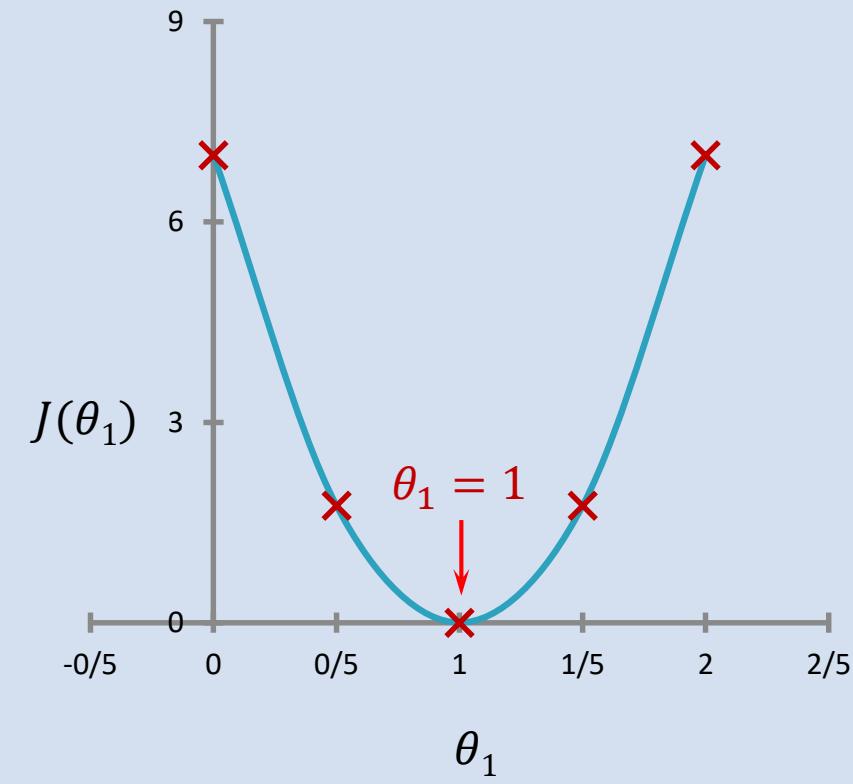
$$J(0) = \frac{1}{2}(1.0^2 + 2.0^2 + 3.0^2) = \frac{1}{2}(14) = 7.0$$

$$J(0) = 7.0$$

تابع هزینه

۱۷

$$\underset{\theta_1}{\text{minimize}} \quad J(\theta_1)$$



اگریوں فٹی تک متغیرہ

۱۸

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

فرضیہ. □

$$\theta_0, \theta_1$$

پارامترها. □

$$J(\theta_0, \theta_1) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

تابع هزینه. □

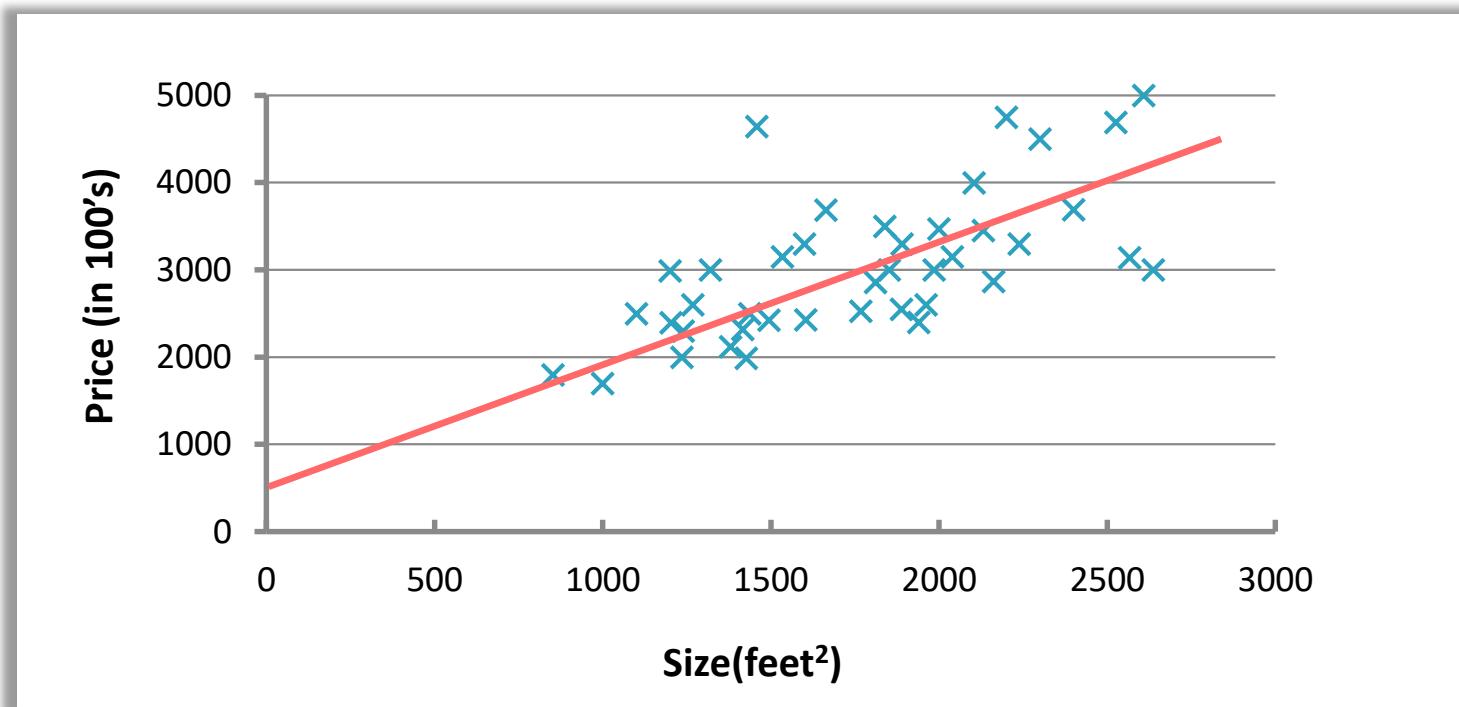
$$\underset{\theta_0, \theta_1}{\text{minimize}} \ J(\theta_0, \theta_1)$$

هدف. □

مثال: قیمت‌گذاری فانه

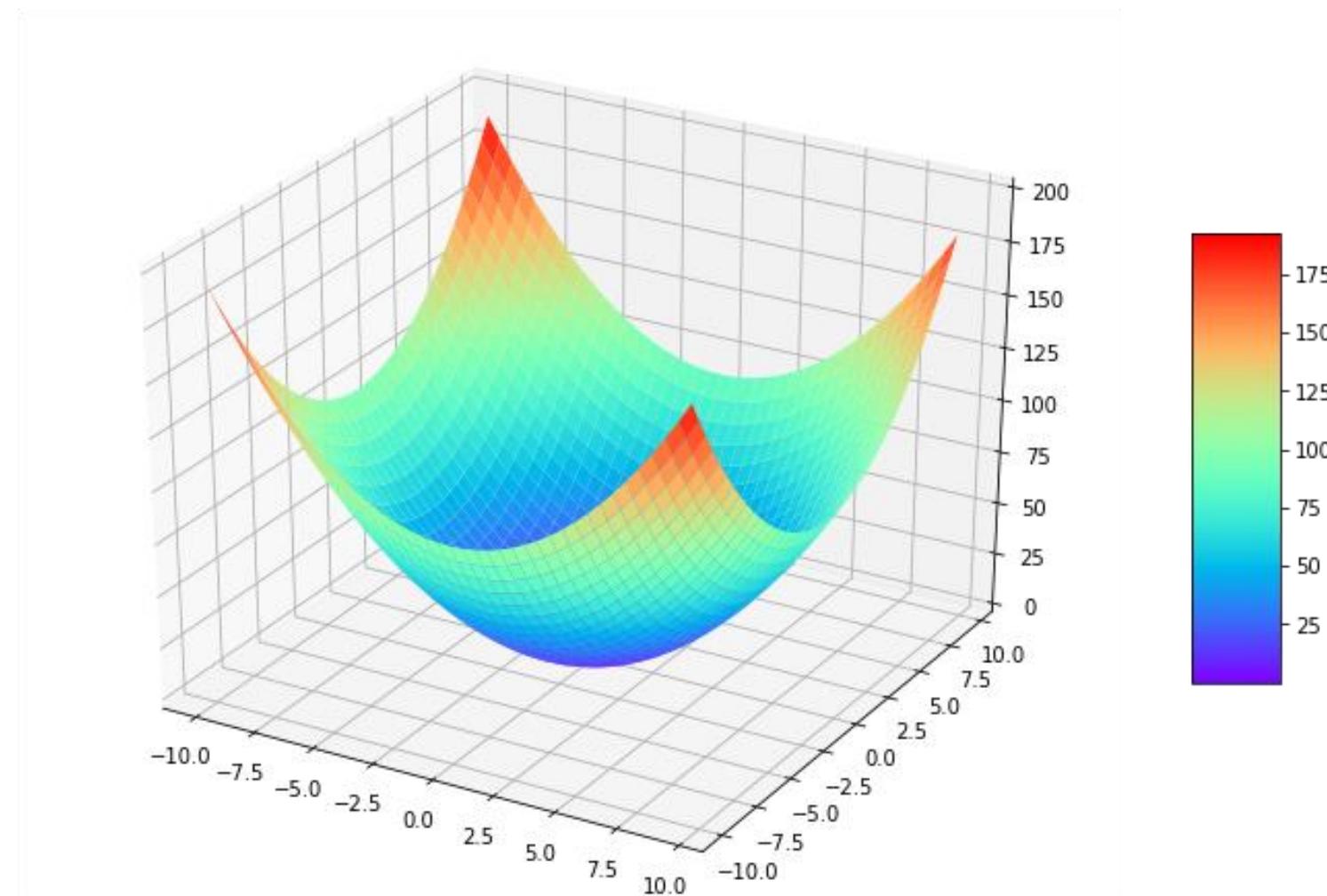
۱۹

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$



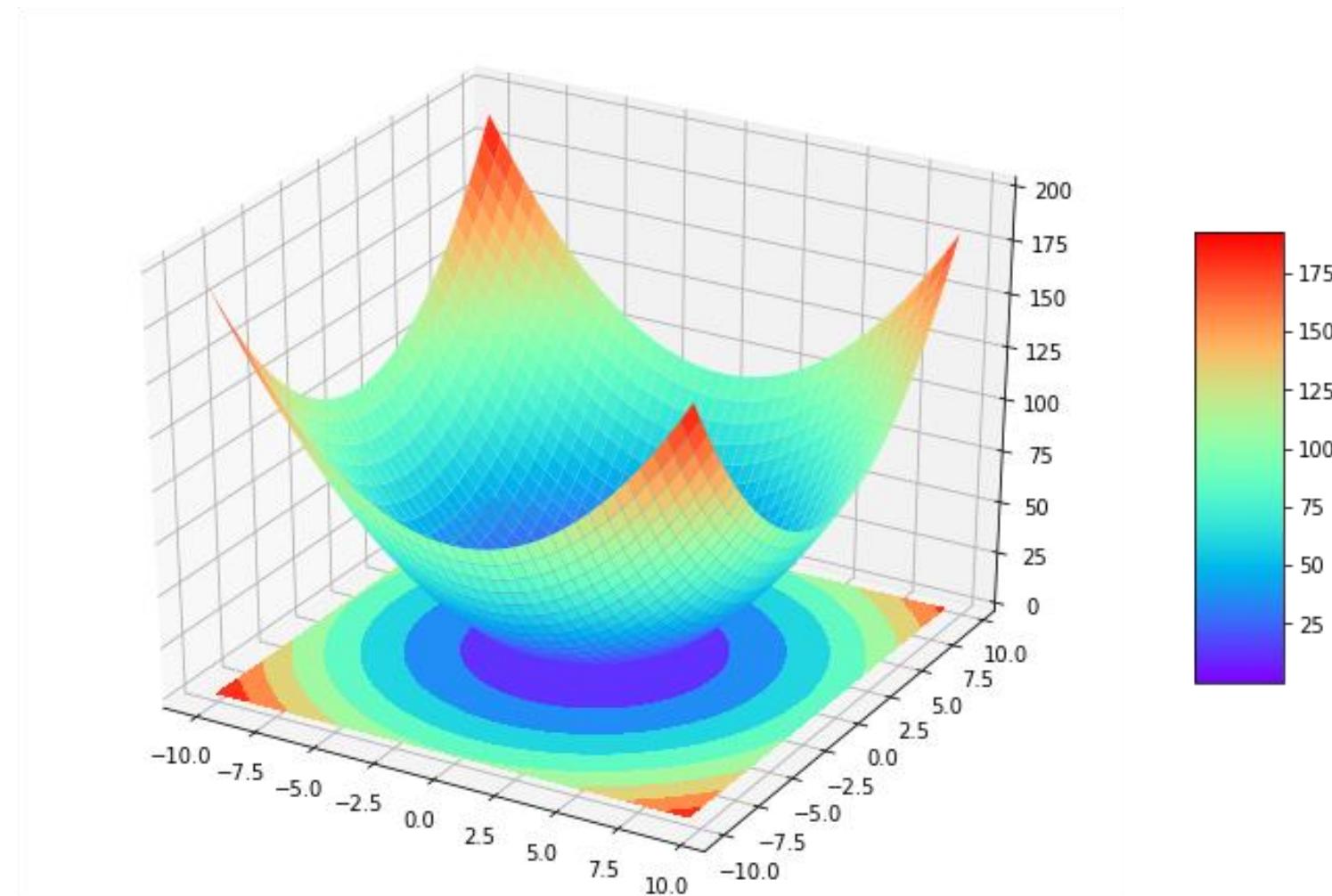
تابع هزینه

۲۰



تابع هزینه

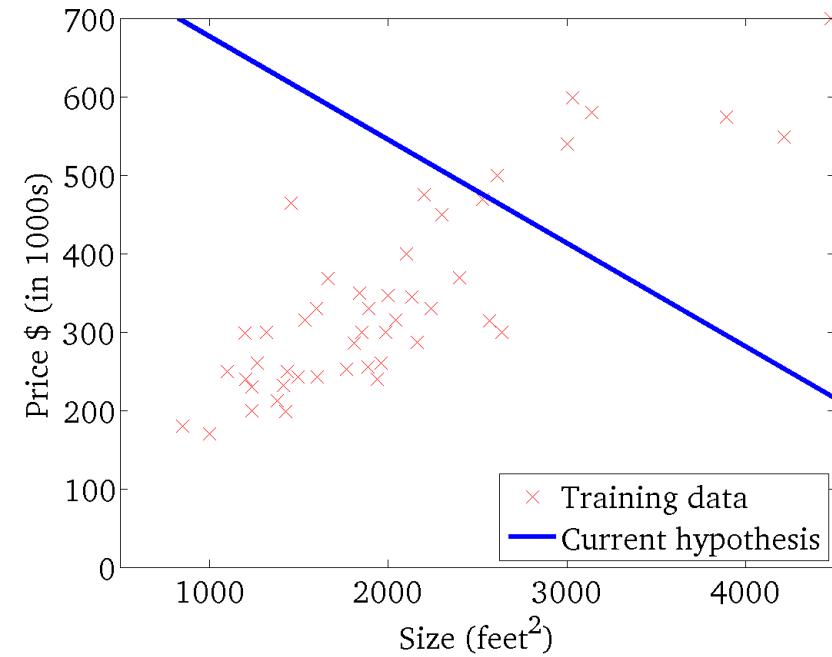
۲۱



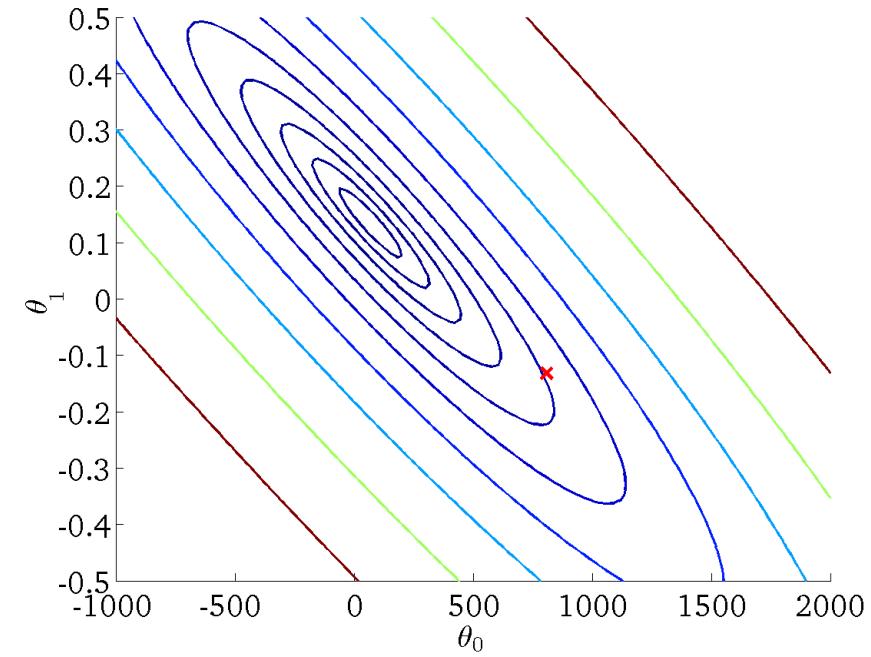
تابع هزینه: متدی کانتور

۲۲

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$



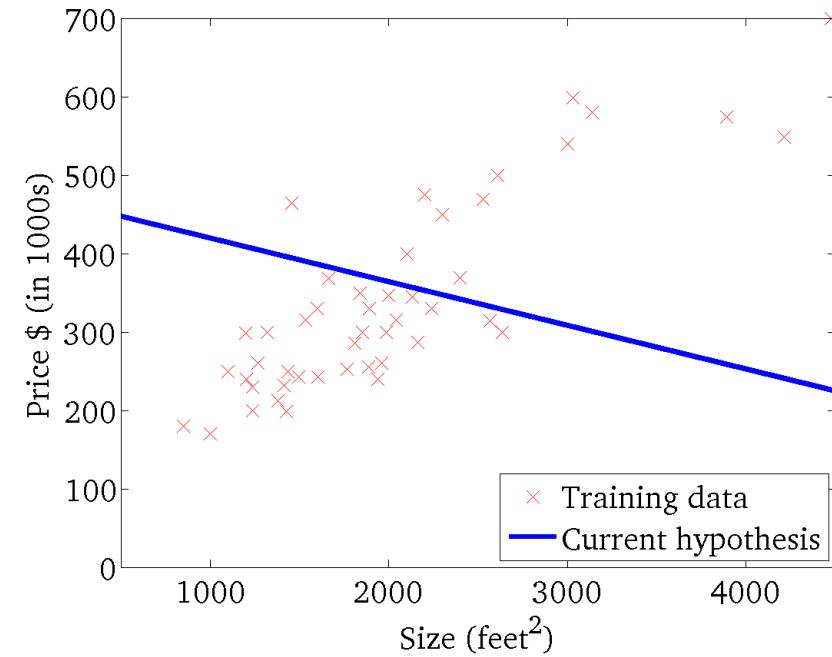
$$J(\theta_0, \theta_1)$$



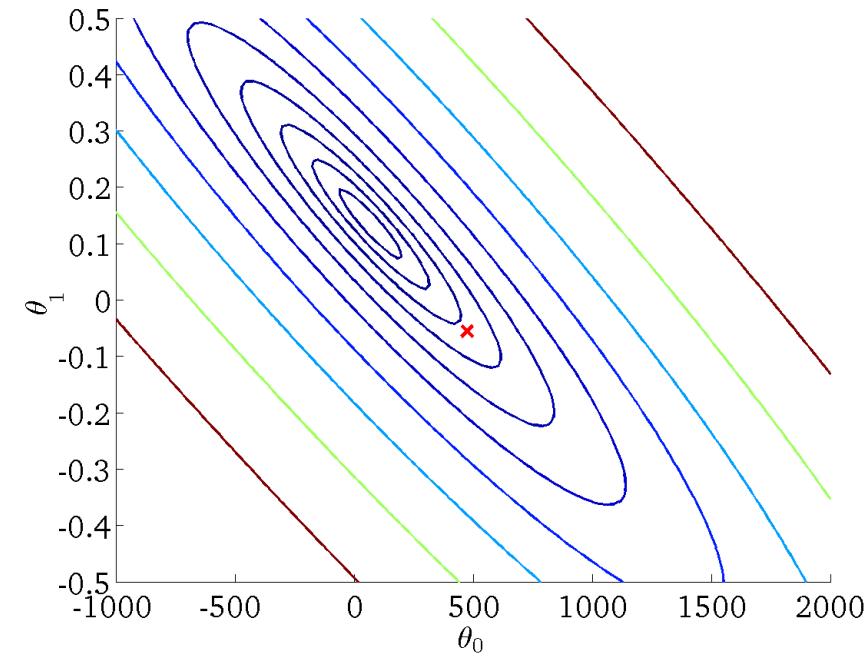
تابع هزینه: متدی کانتور

۲۳

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$



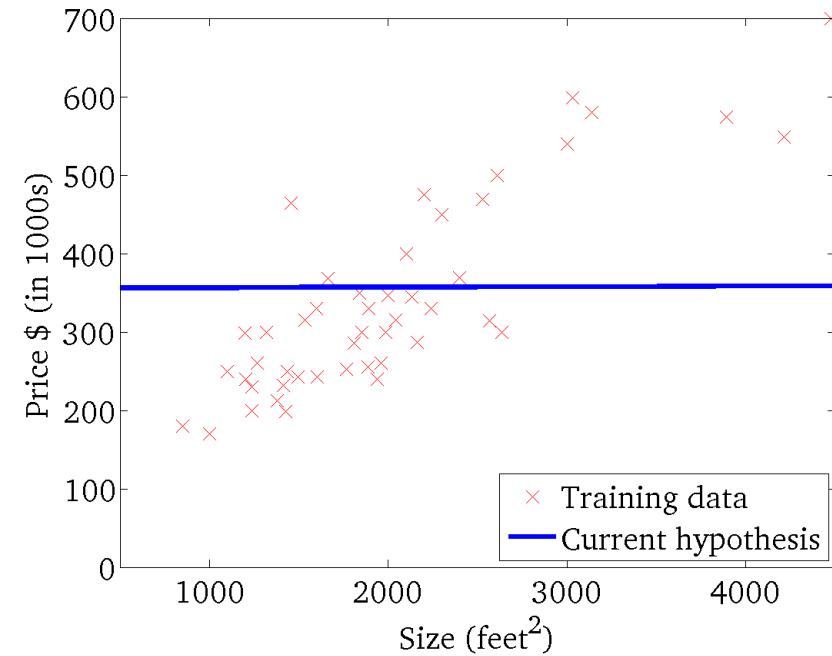
$$J(\theta_0, \theta_1)$$



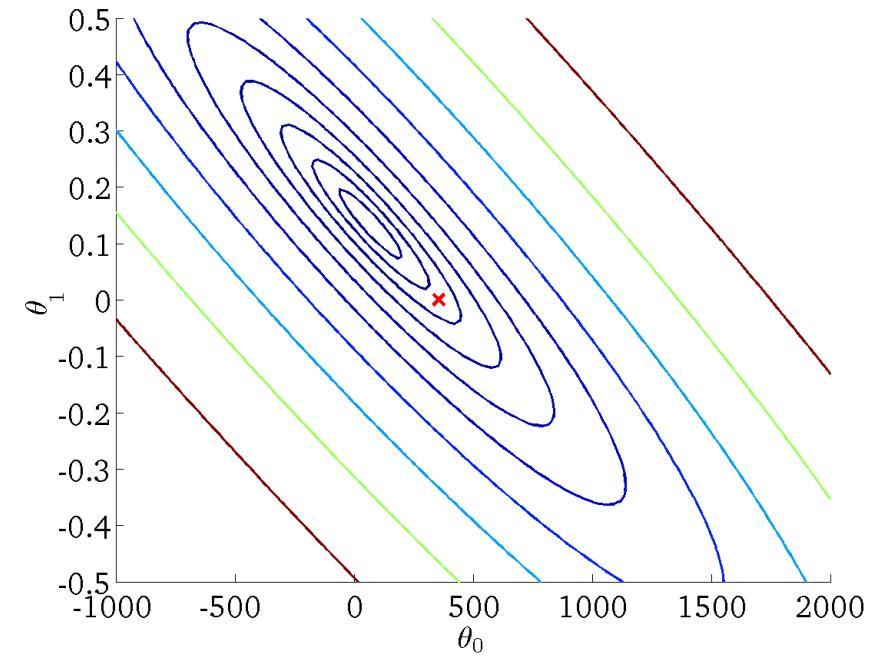
تابع هزینه: متدی کانتور

۲۴

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$



$$J(\theta_0, \theta_1)$$

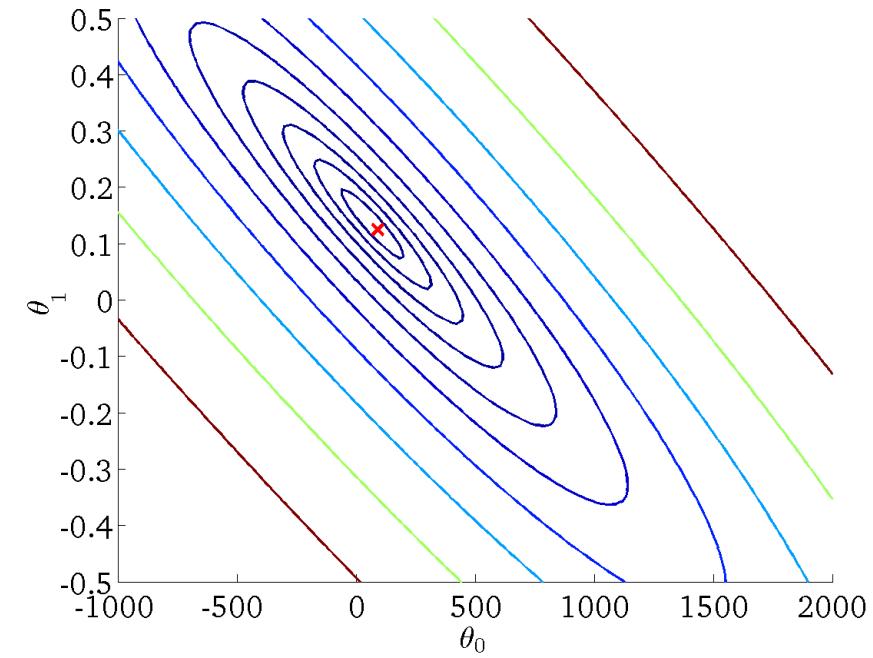
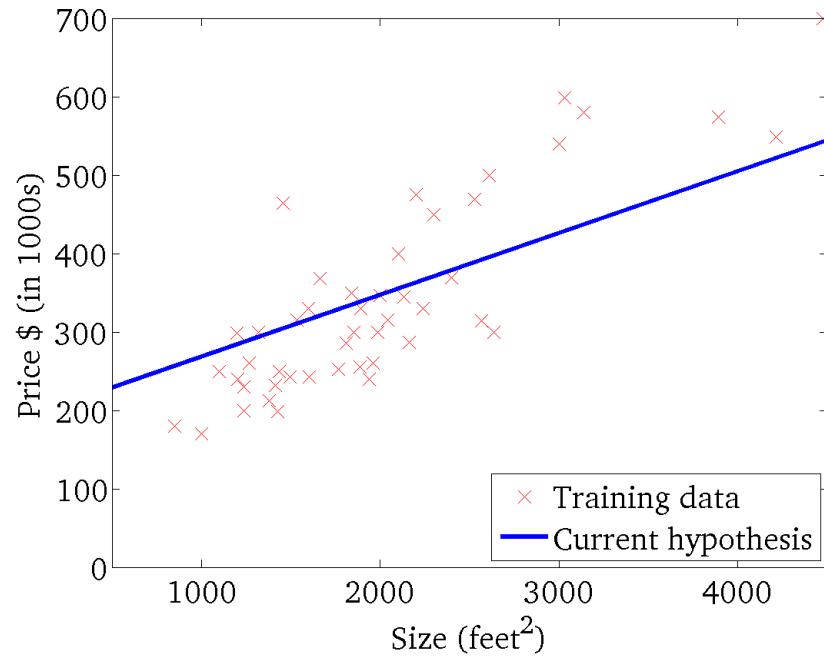


تابع هزینه: متدی کانتور

۲۵

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta_0, \theta_1)$$



گرایان گاہشی

□ تابع هزینه.

$$J(\theta_0, \theta_1)$$

□ هدف.

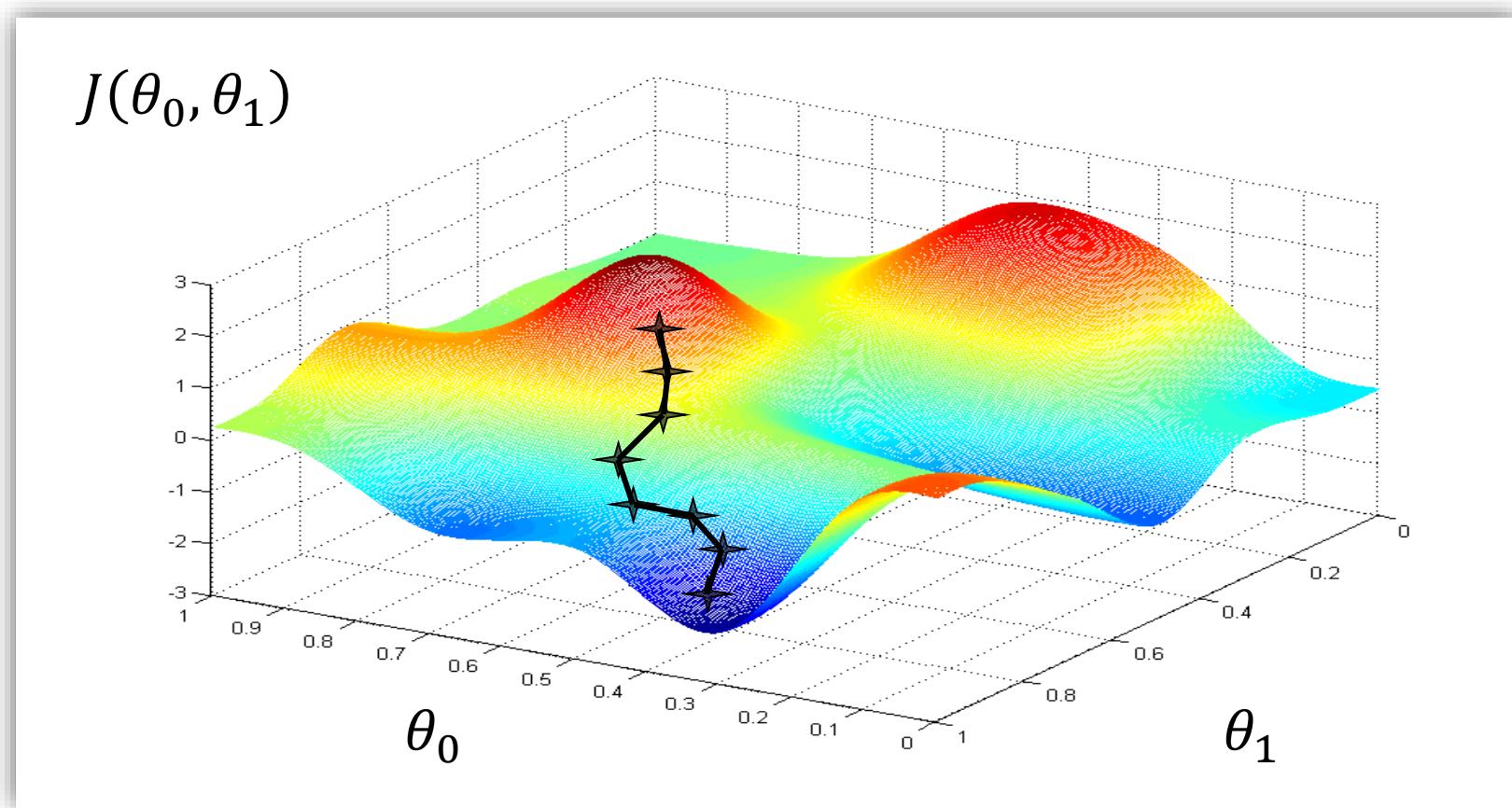
$$\underset{\theta_0, \theta_1}{\text{minimize}} \ J(\theta_0, \theta_1)$$

□ کلیات روش.

- با یک مقدار اولیه تصادفی برای پارامترهای θ_0 و θ_1 شروع کن. [مثالاً مقدار صفر]
- مقدار پارامترها را به گونه‌ای تغییر بده که مقدار تابع هزینه $J(\theta_0, \theta_1)$ کاهش یابد.
- عمل بالا را آن قدر تکرار کن تا به یک مقدار کمینه برای تابع هزینه برسیم. [همگرایی]

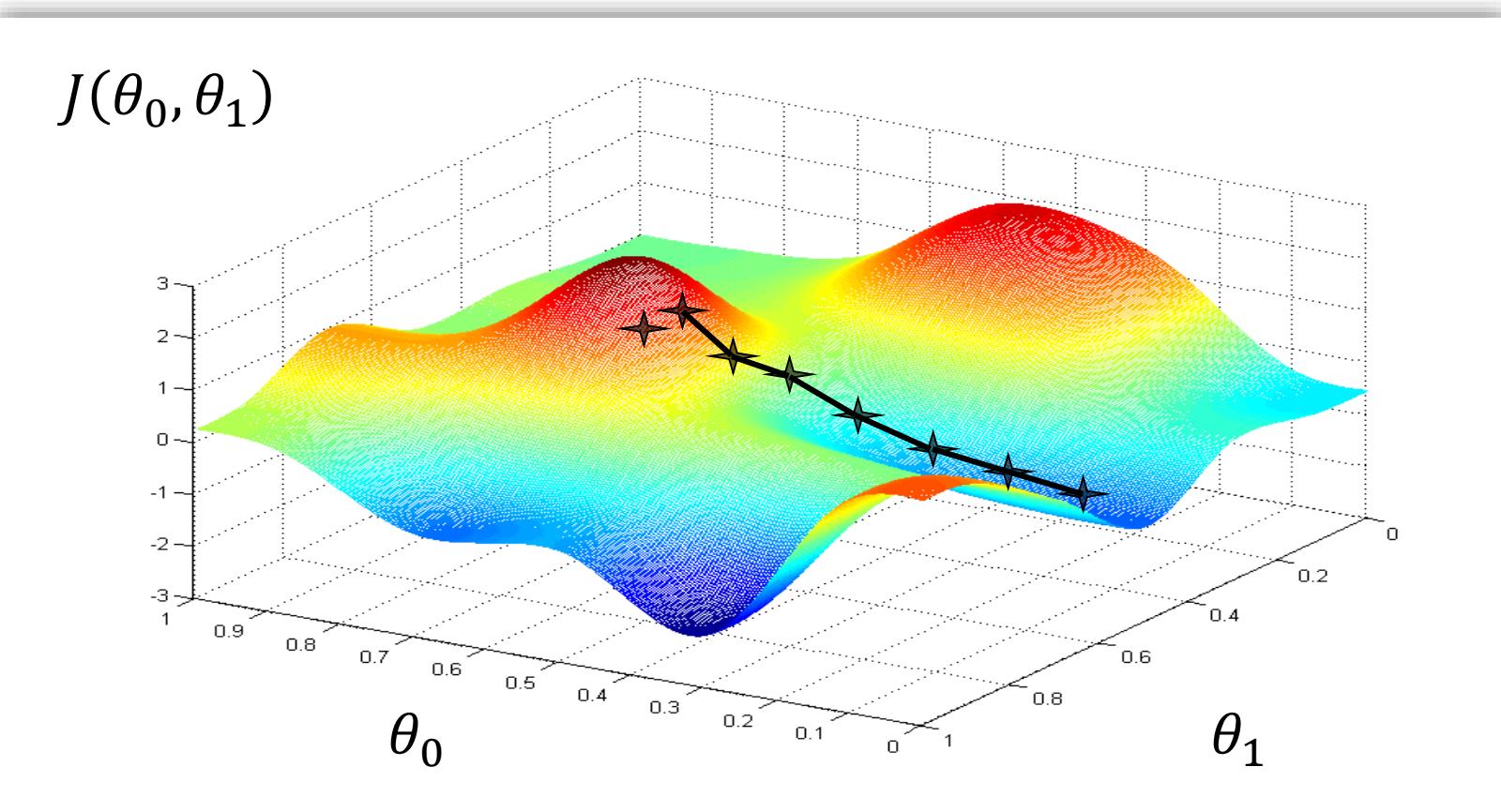
گرادیان کاہشی: بهینه سراسری

۲۸



گرادیان کاہشی: بهینه محلی

۲۹



الگوریتم گرادیان کاہشی

۳۰

repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad (\text{for } j = 0 \text{ and } j = 1)$$

}

نحو یادگیری

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

□ پیاده‌سازی درست. به روزرسانی مقدار پارامترها به طور همزمان

$$\Delta\theta_0 := -\alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$

$$\Delta\theta_1 := -\alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$

$$\theta_0 := \theta_0 + \Delta\theta_0$$

$$\theta_1 := \theta_1 + \Delta\theta_1$$

الگوریتم گرادیان کاہشی

۳۱

repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad (\text{for } j = 0 \text{ and } j = 1)$$

}

نخ یادگیری

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

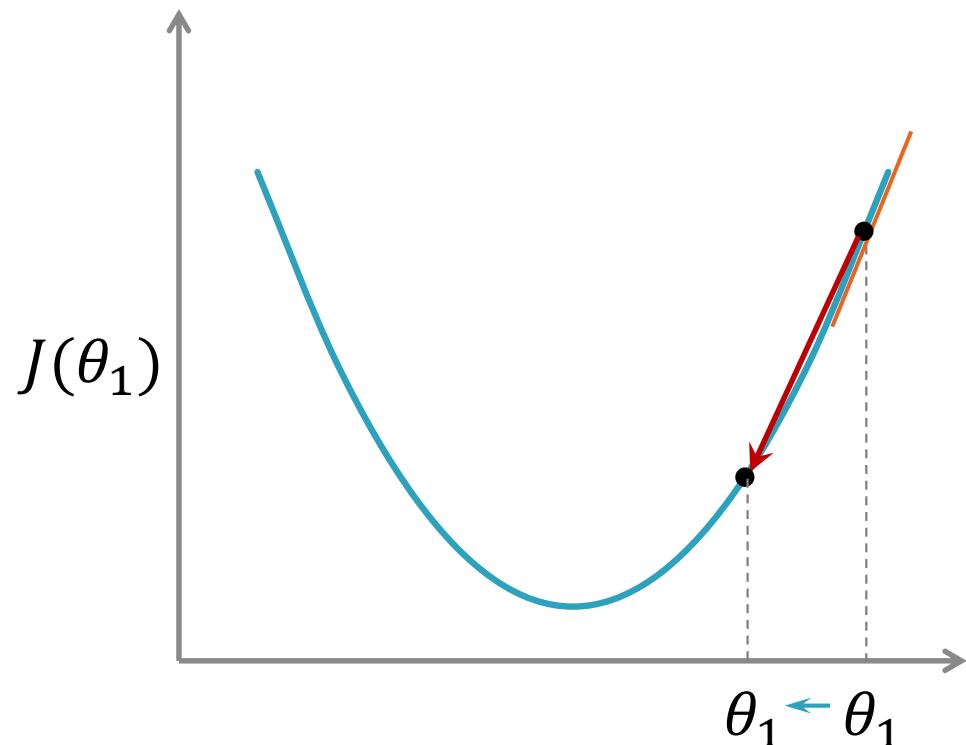
□ پیاده‌سازی نادرست. به روزرسانی مقدار پارامترها به طور ترتیبی

$$\theta_0 := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$

$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$

الگوريتم گراديان کاهشی: گراديان

۳۲



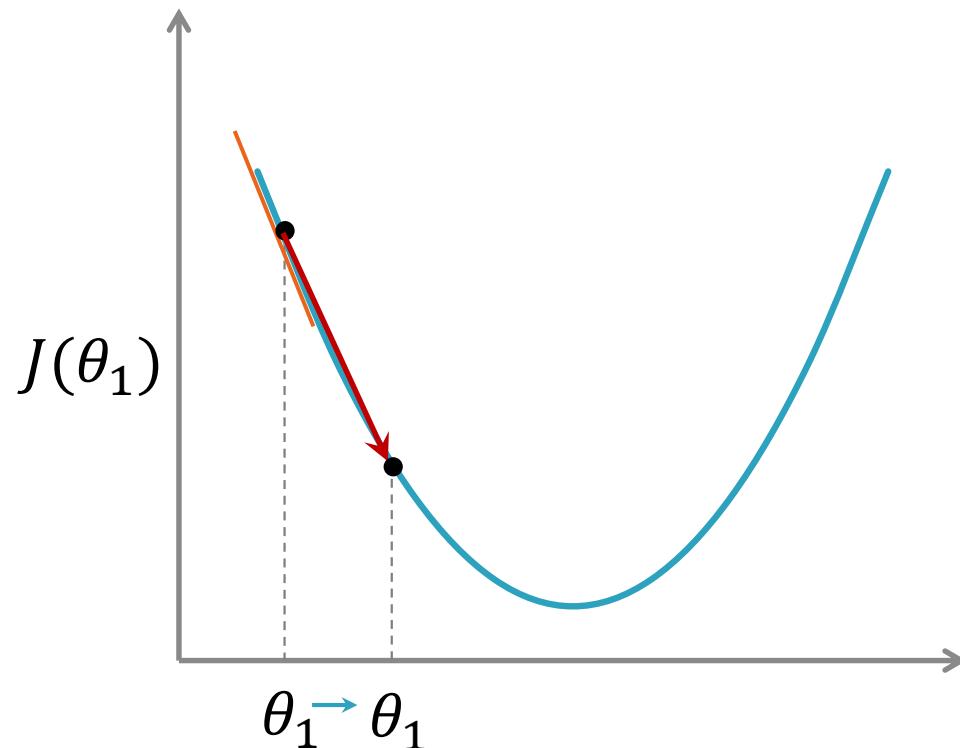
شيب مثبت

$$\theta_1 := \theta_1 - \alpha \boxed{\frac{\partial}{\partial \theta_1} J(\theta_1)}$$

$$\theta_1 := \theta_1 - \underbrace{\alpha}_{\geq 0} (\geq 0)$$

الگوريتم گراديان کاهشی: گراديان

۳۳



شيب منفي

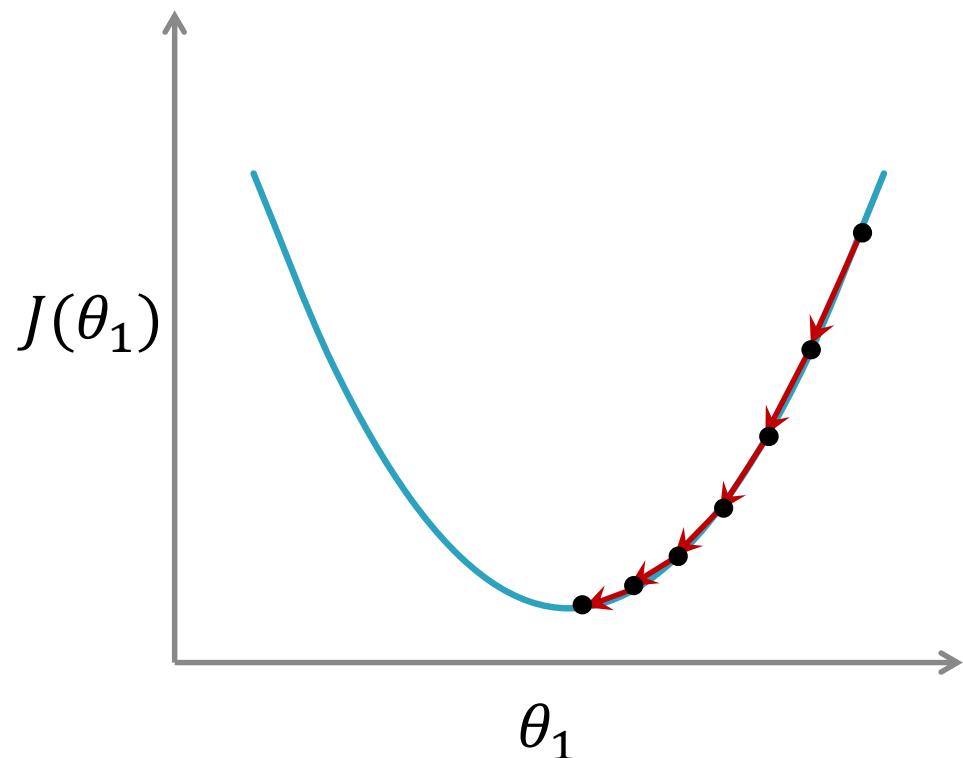
$$\theta_1 := \theta_1 - \alpha \boxed{\frac{\partial}{\partial \theta_1} J(\theta_1)}$$

$$\theta_1 := \theta_1 - \underbrace{\alpha (\leq 0)}_{\leq 0}$$

الگوریتم گرادیان کاهشی: نرخ یادگیری

۳۴

اگر نرخ یادگیری بیش از حد کوچک باشد، گرادیان کاهشی به کندی همگرا خواهد شد. □

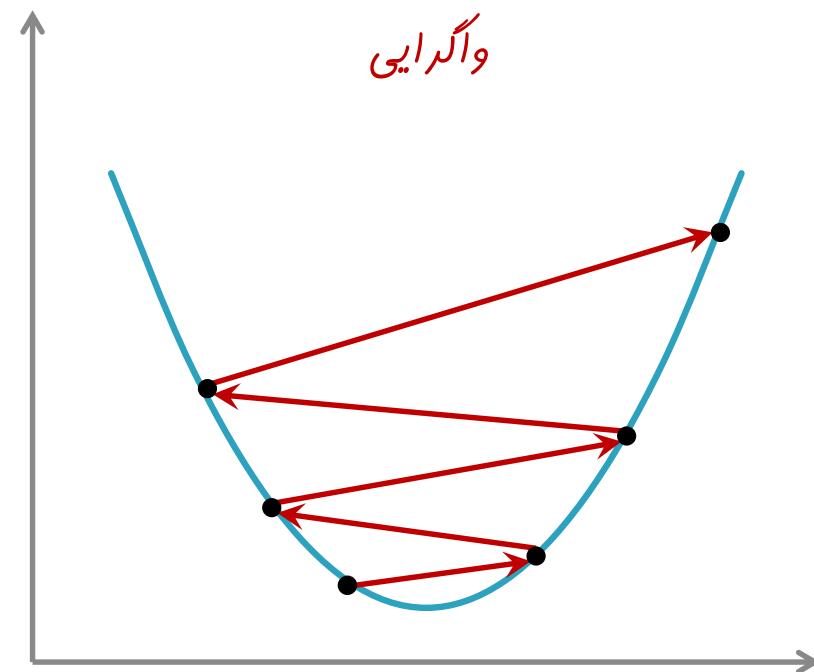
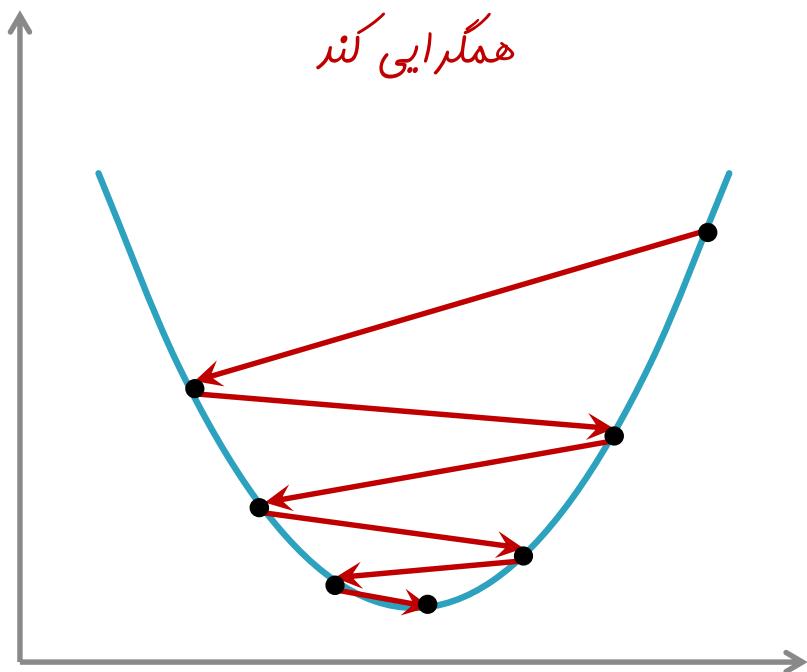


$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

الگوریتم گرادیان کاہشی: نرخ یادگیری

۳۵

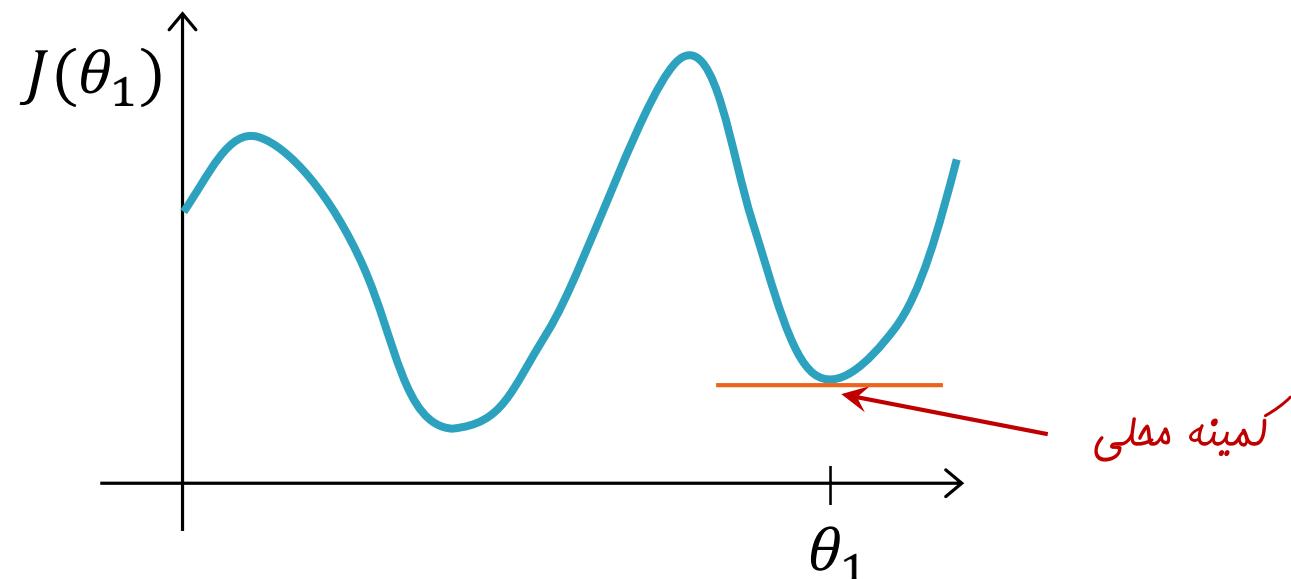
□ اگر نرخ یادگیری بیش از حد بزرگ باشد، گرادیان کاہشی ممکن است به کندی همگرا شود و یا حتی واگرا شود.



الگوریتم گرادیان کاہشی: همگرایی

۳۶

□ همگرایی: زمانی که مقدار پارامتر θ_1 در یک کمینه محلی قرار بگیرد.



$$\theta_1 := \theta_1 - \alpha \underbrace{\frac{\partial}{\partial \theta_1} J(\theta_1)}_{صفر}$$

کاربرد گرادیان کاہشی در رگرسیون خطی

گرادیان کاہشی و رگرسیون خطی

۳۸

repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad (\text{for } j = 0 \text{ and } j = 1)$$

}

□ رگرسیون خطی.



$$J(\theta_0, \theta_1) = \frac{1}{2} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

$$h_\theta(x) = \theta_0 + \theta_1 x$$

گرادیان تابع هزینه

۳۹

$$\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) = \frac{\partial}{\partial \theta_j} \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 = \frac{\partial}{\partial \theta_j} \frac{1}{2} \sum_{i=1}^m ((\theta_0 + \theta_1 x^{(i)}) - y^{(i)})^2$$

$$j = 0 \Rightarrow \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})$$

$$j = 1 \Rightarrow \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) \cdot x^{(i)}$$

گرادیان کاہشی و اگرسیون خطي

۴۰

repeat until convergence {

$$\theta_0 := \theta_0 - \alpha \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})$$

$$\theta_1 := \theta_1 - \alpha \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x^{(i)}$$

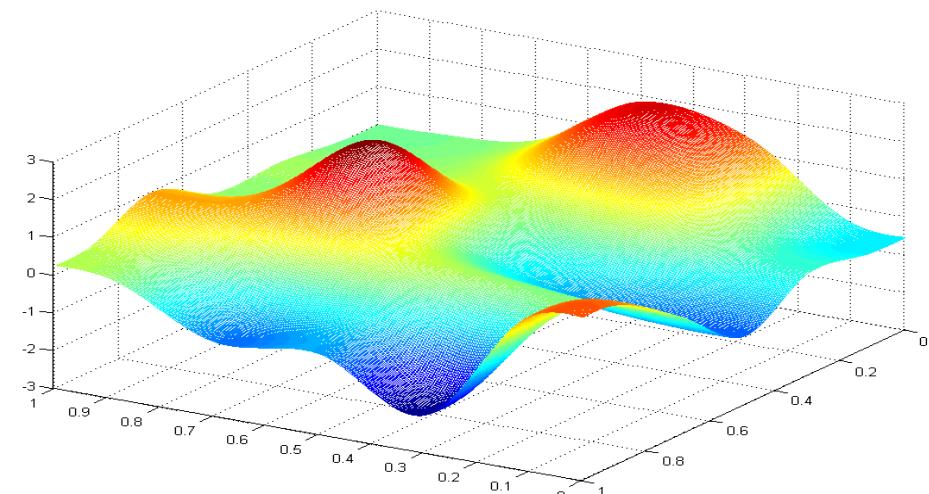
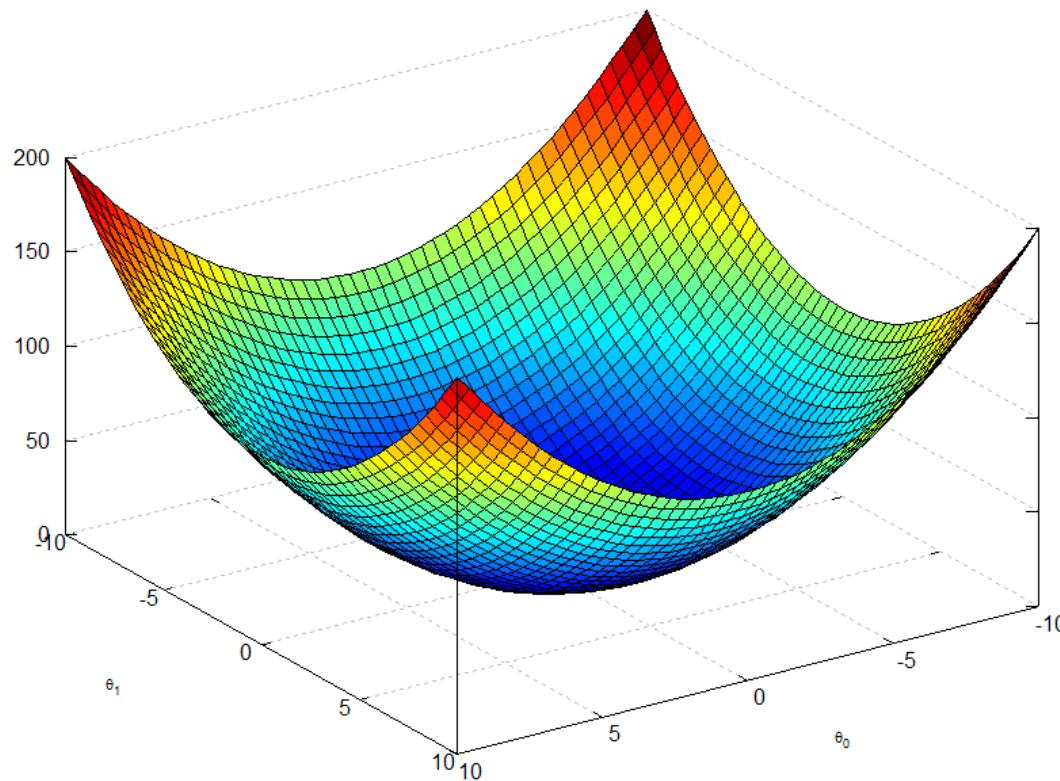
}

به روزرسانی همزمان

گرادیان کاهشی و رگرسیون خطی

۴۱

□ توجه. در رگرسیون خطی تابع هزینه یک **تابع کوثر** است و در نتیجه گرادیان کاهشی در صورت همگرایی لزوماً در بهینه‌ی سراسری همگرا می‌شود.

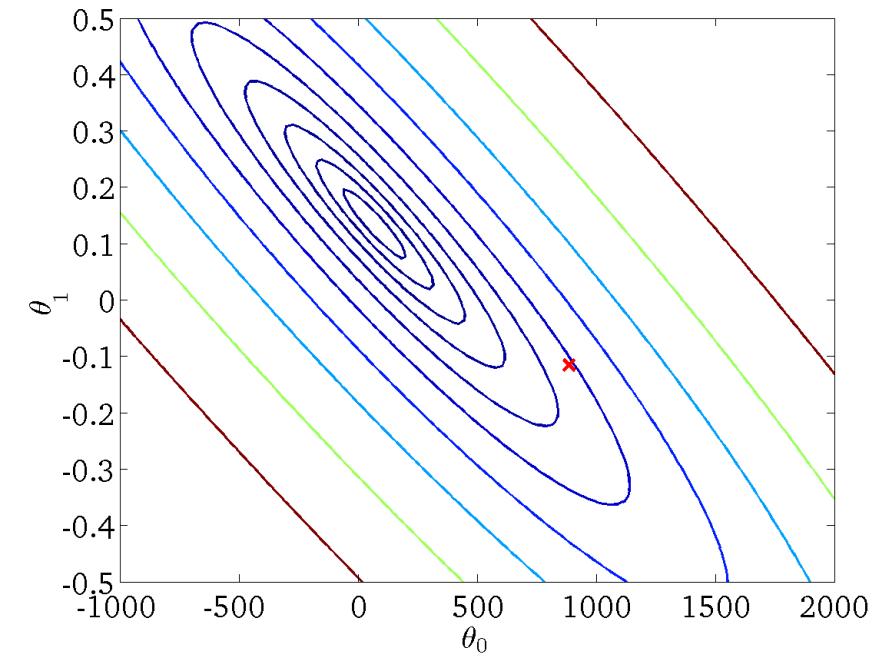
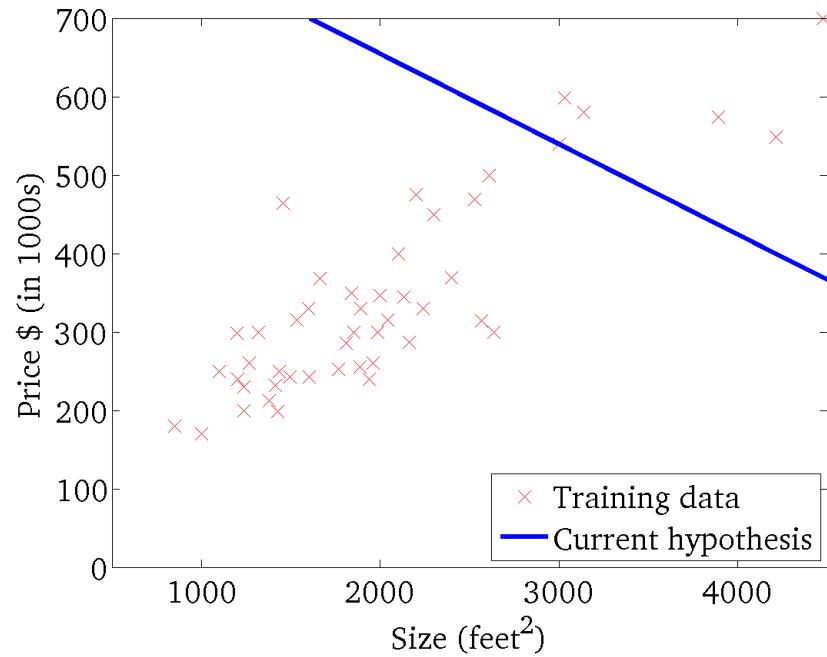


گرادیان کاہشی

۴۲

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta_0, \theta_1)$$

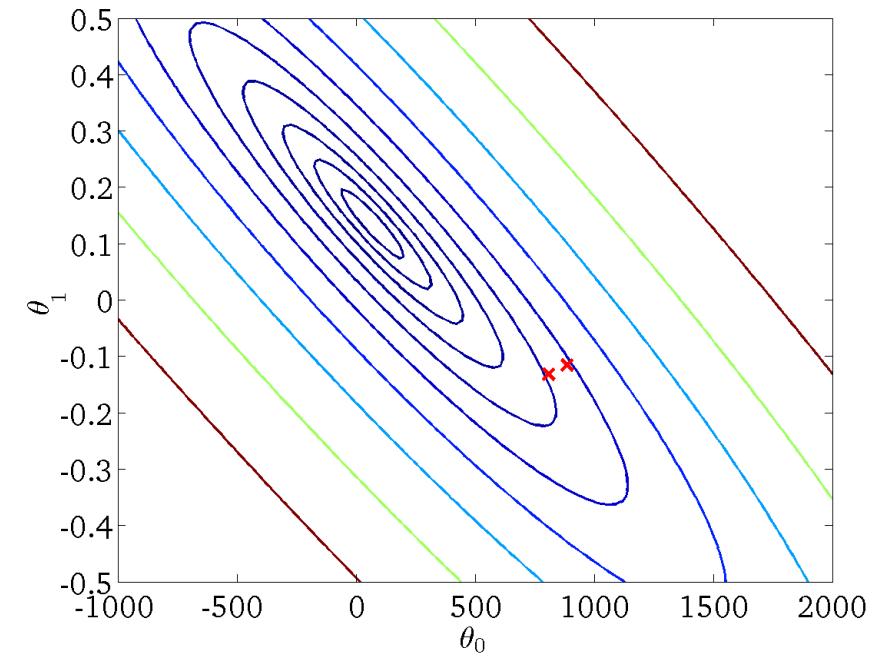
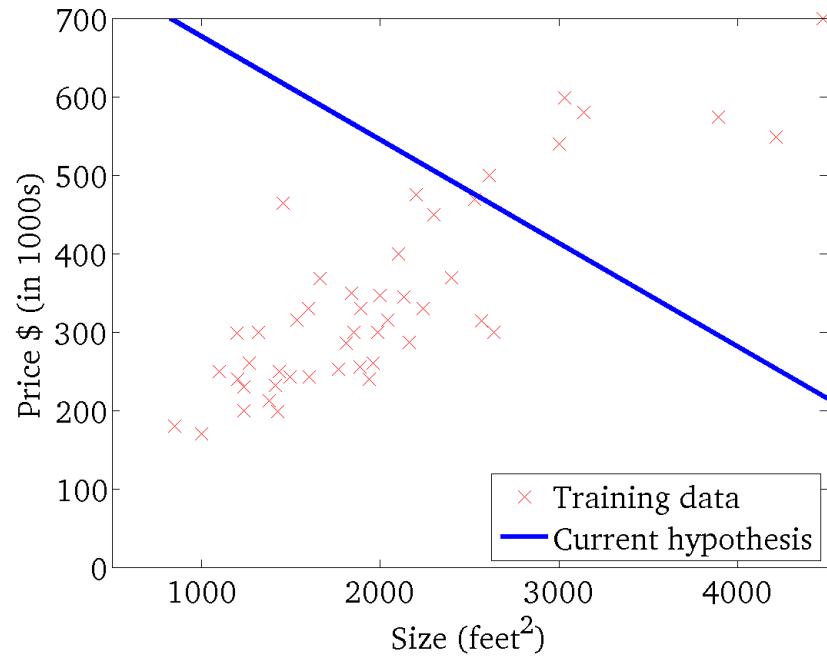


گرادیان کاہشی

۴۳

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta_0, \theta_1)$$

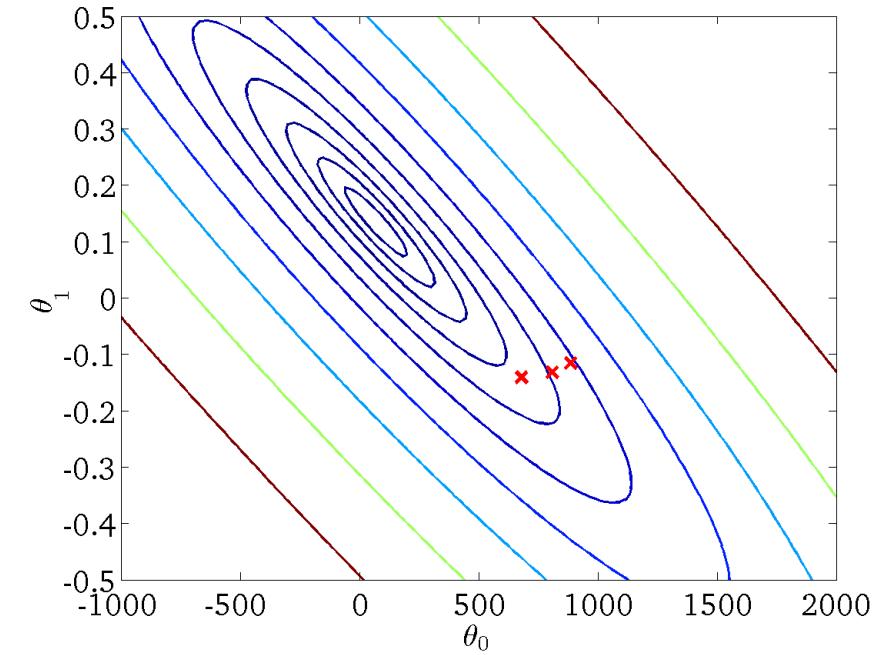
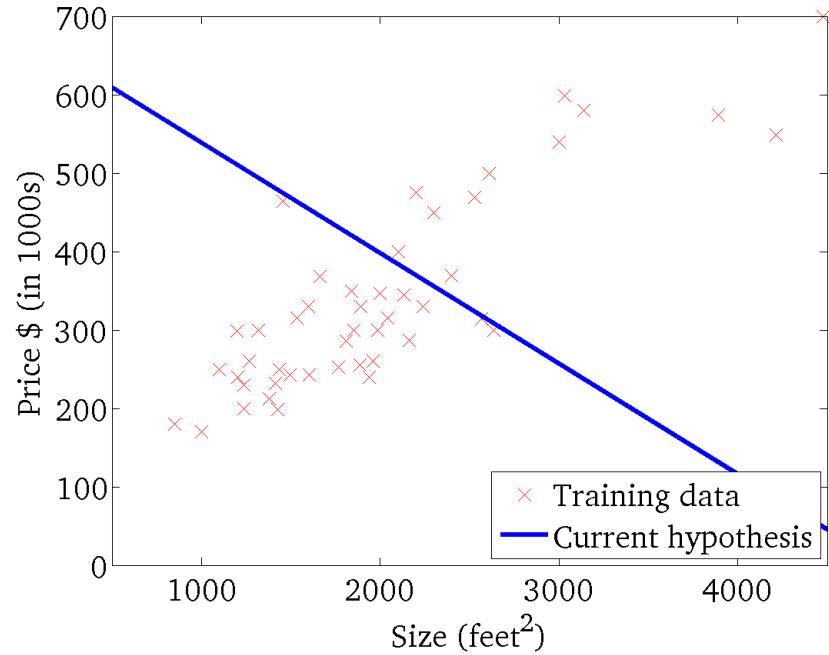


گرادیان کاہشی

۴۴

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta_0, \theta_1)$$

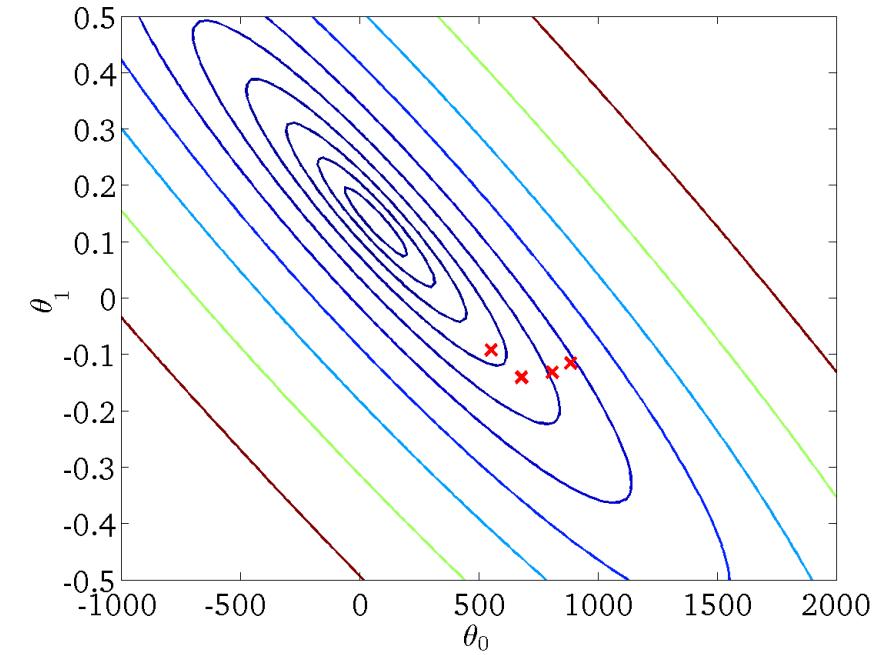
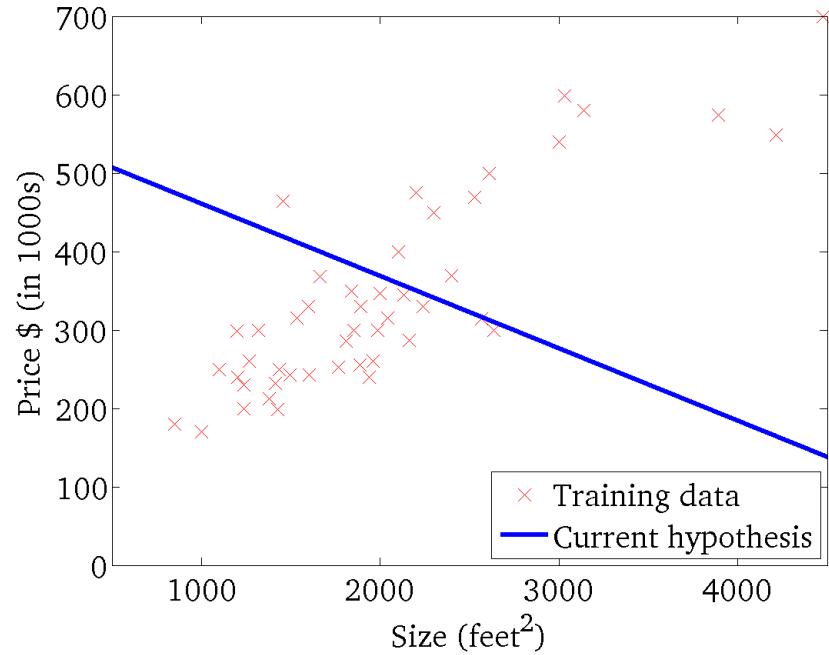


گرادیان کاہشی

۴۵

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta_0, \theta_1)$$

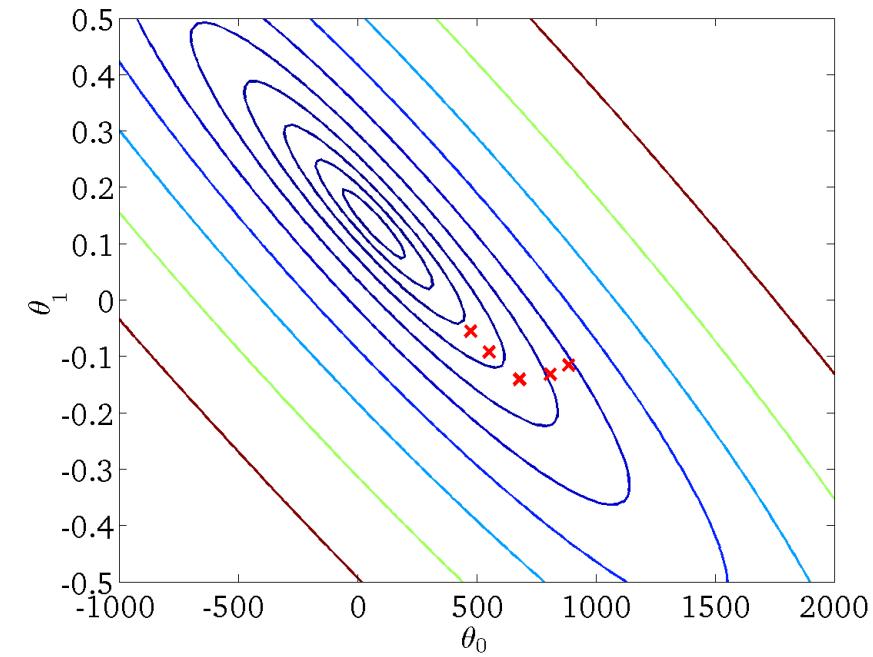
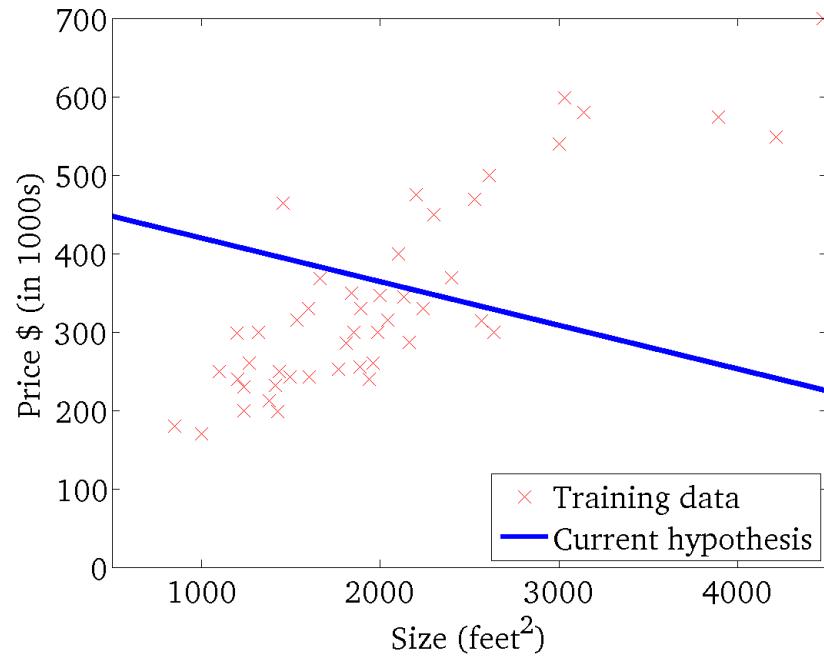


گرادیان کاہشی

۴۶

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta_0, \theta_1)$$

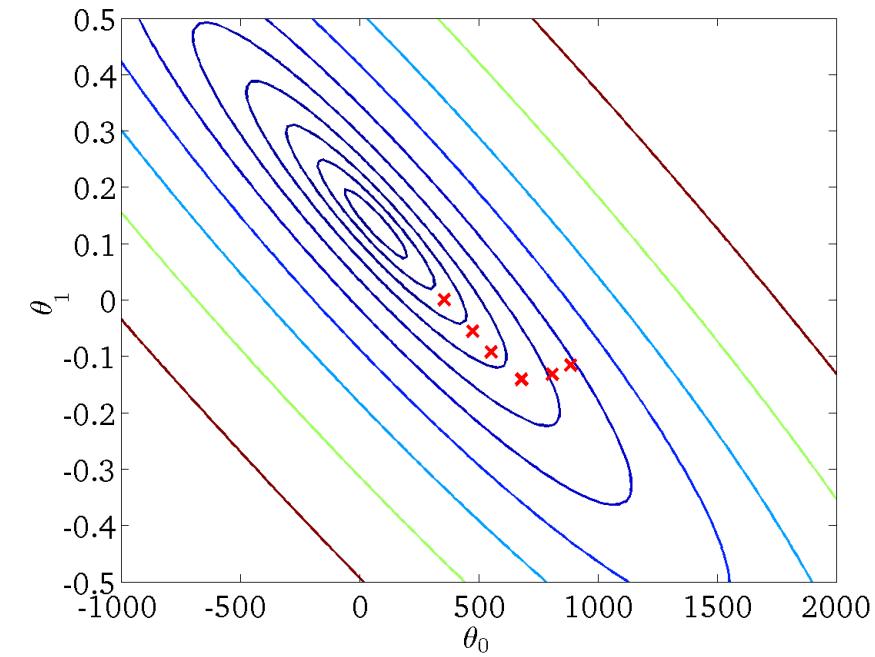
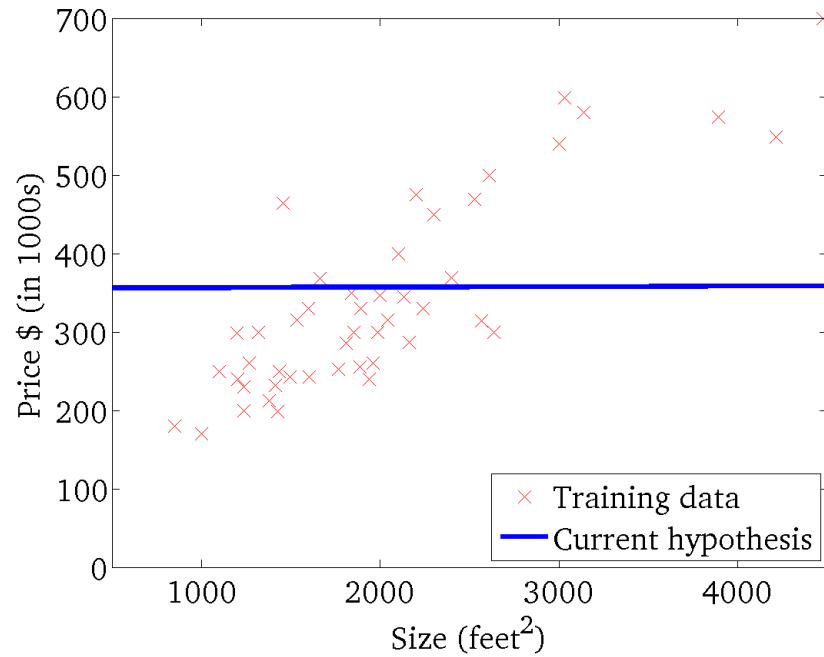


گرادیان کاہشی

۴۷

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta_0, \theta_1)$$

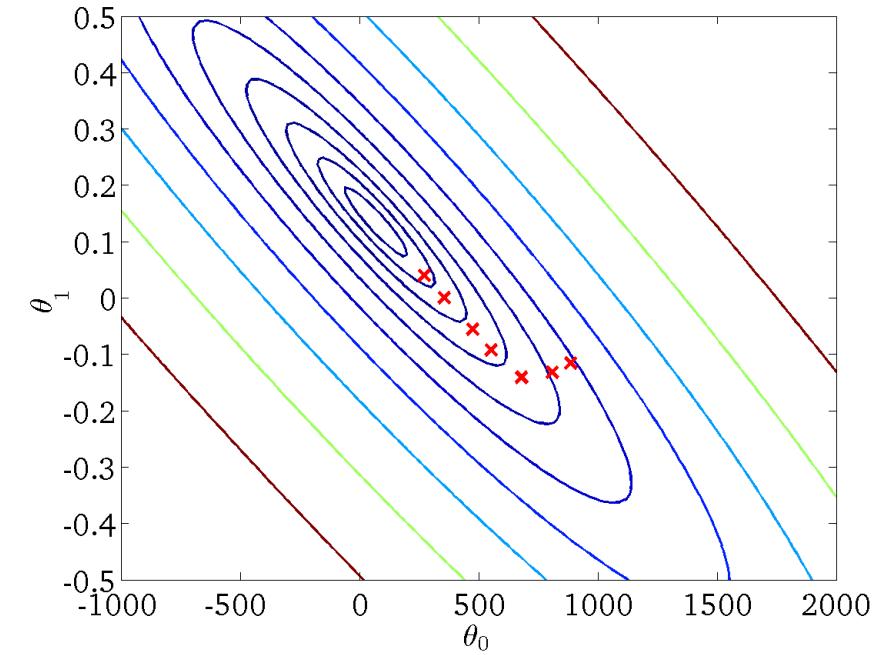
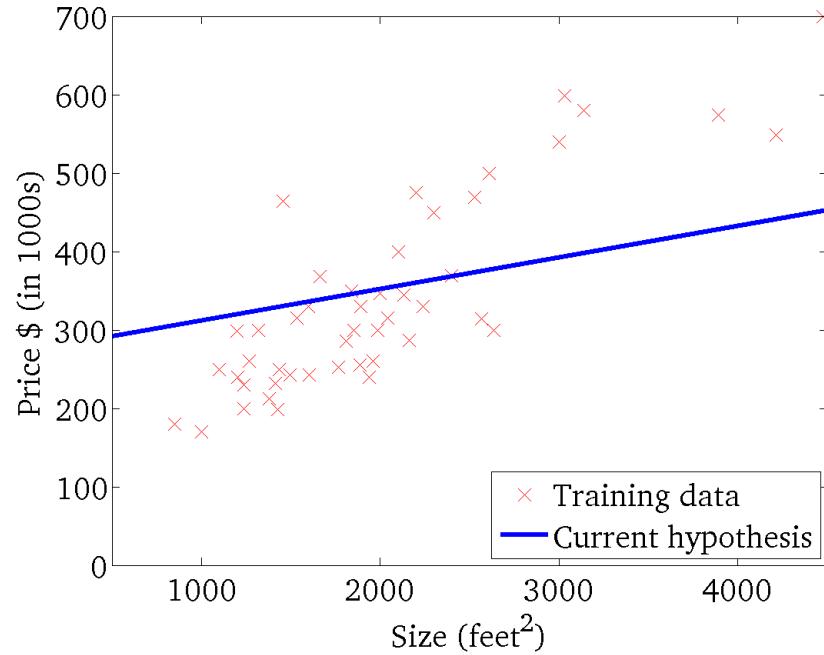


گرادیان کاہشی

۴۸

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta_0, \theta_1)$$

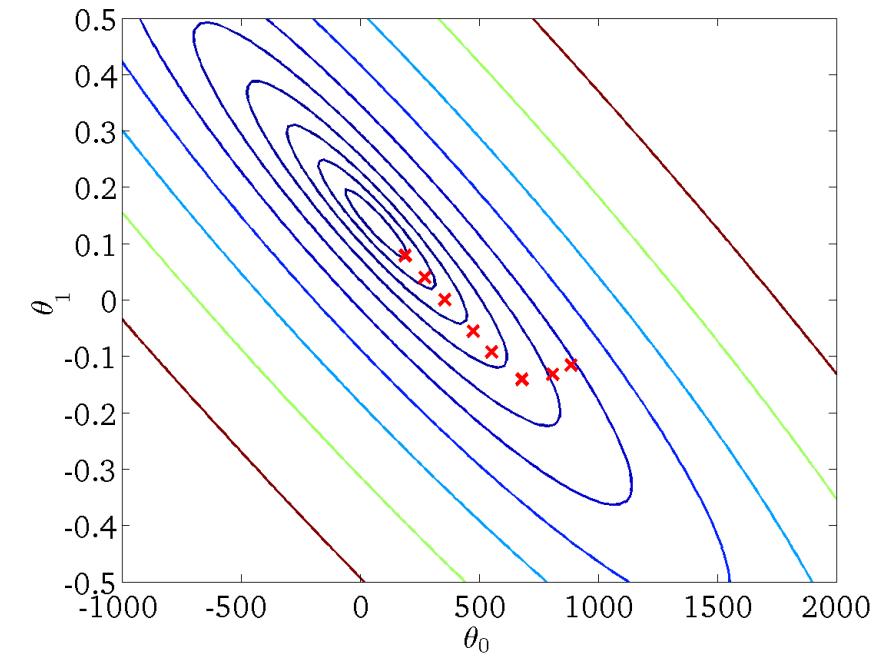
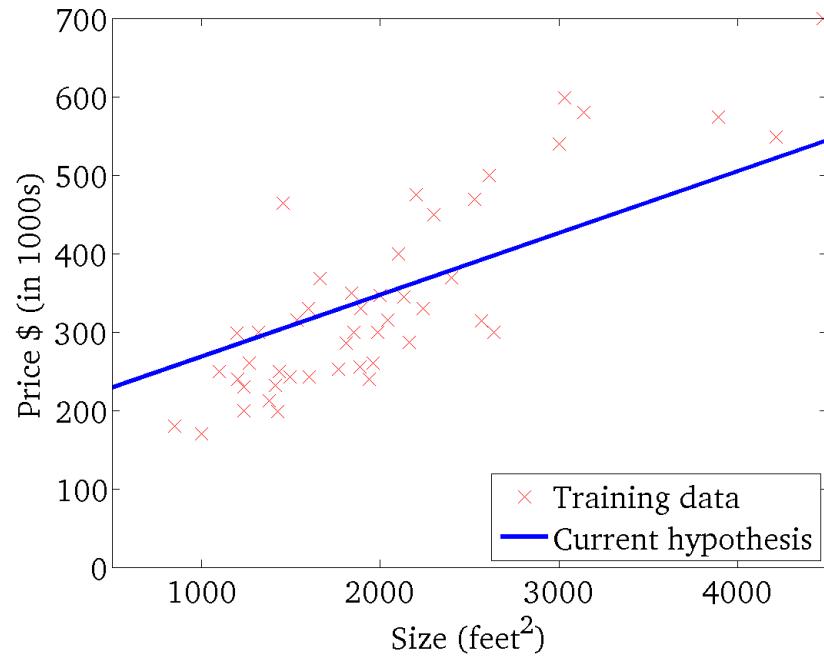


گرادیان کاہشی

۴۹

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta_0, \theta_1)$$

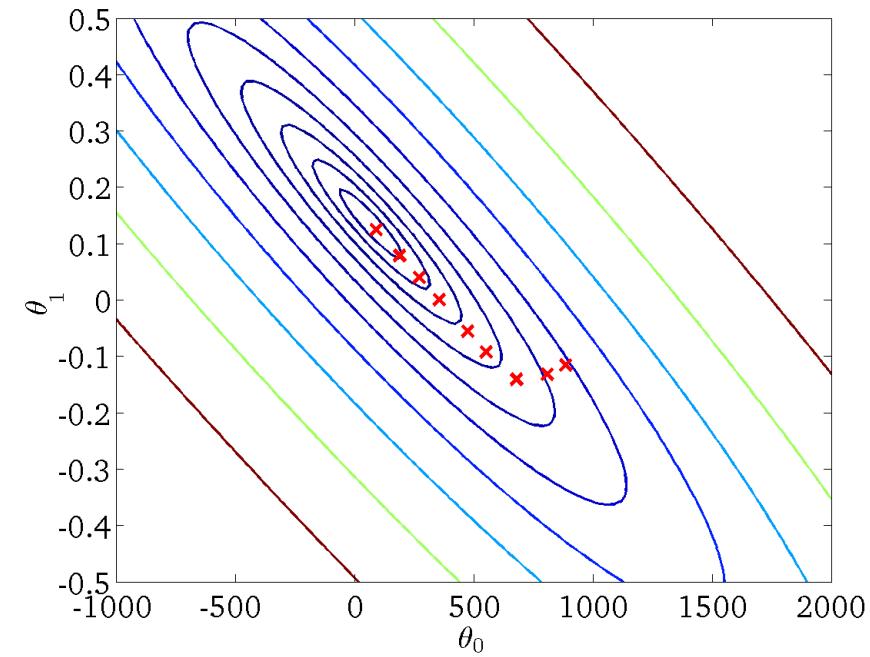
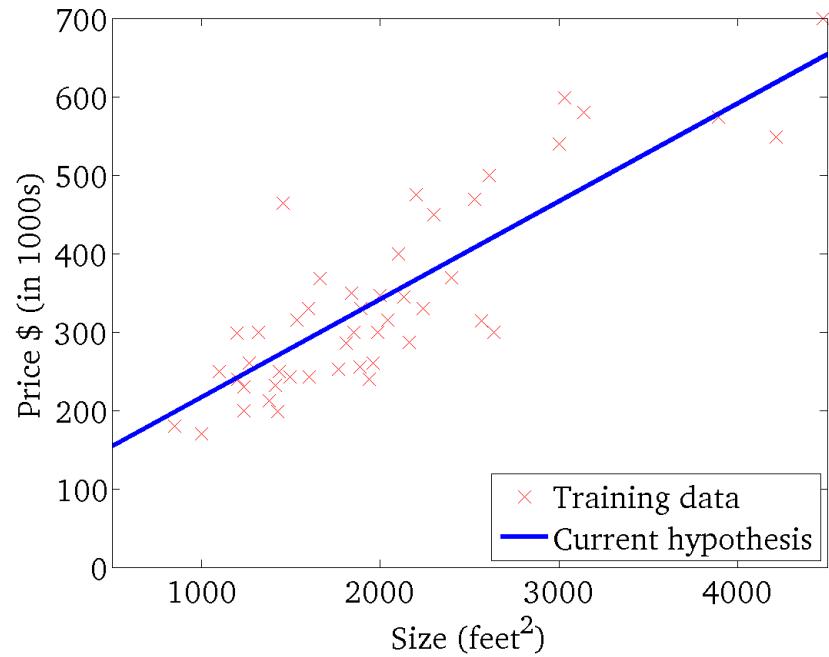


گرادیان کاہشی

۵.

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

$$J(\theta_0, \theta_1)$$



گرادیان کاهشی دسته‌ای

۵۱

□ گرادیان کاهشی دسته‌ای. در هر تکرار الگوریتم، از تمام نمونه‌های آموزشی برای به روز رسانی مقدار پارامترها استفاده می‌شود.

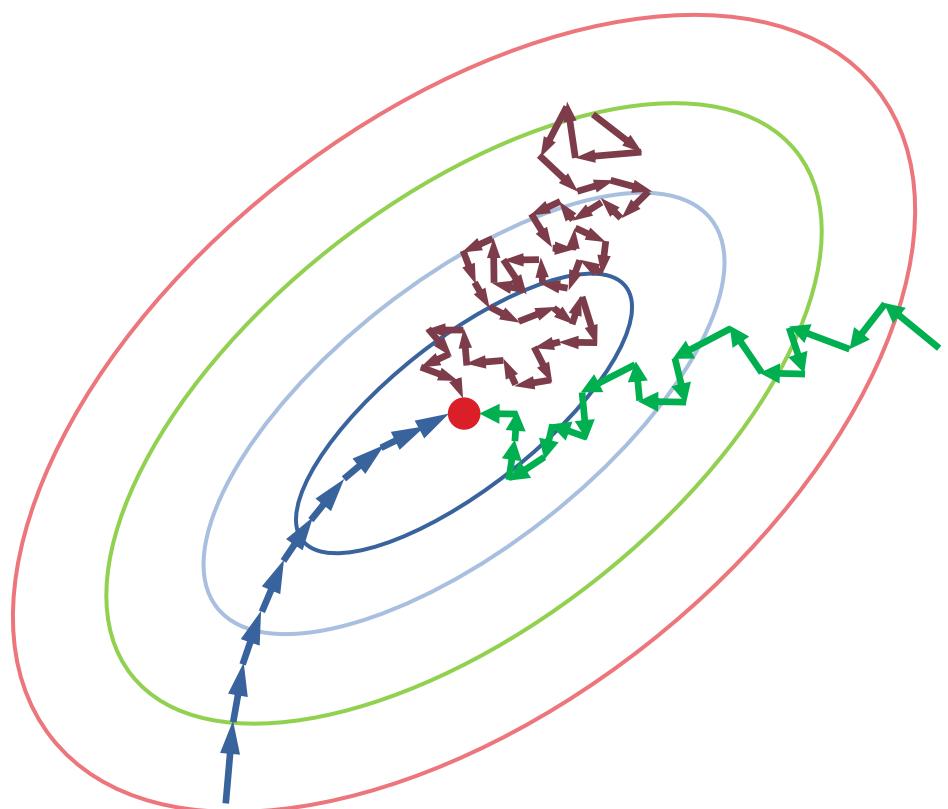
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \quad (\text{for } j = 0 \text{ and } j = 1)$$

$$j = 0 \Rightarrow \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})$$

$$j = 1 \Rightarrow \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x^{(i)}$$

انواع گرادیان کاهشی

۵۲



—
گرادیان کاهشی با دسته‌های کامل

به روز رسانی در هر تکرار با استفاده از تمام نمونه‌ها

—
گرادیان کاهشی اتفاقی (آنلاین)

به روز رسانی در هر تکرار با استفاده از یک نمونه تصادفی

—
گرادیان کاهشی با دسته‌های کوچک

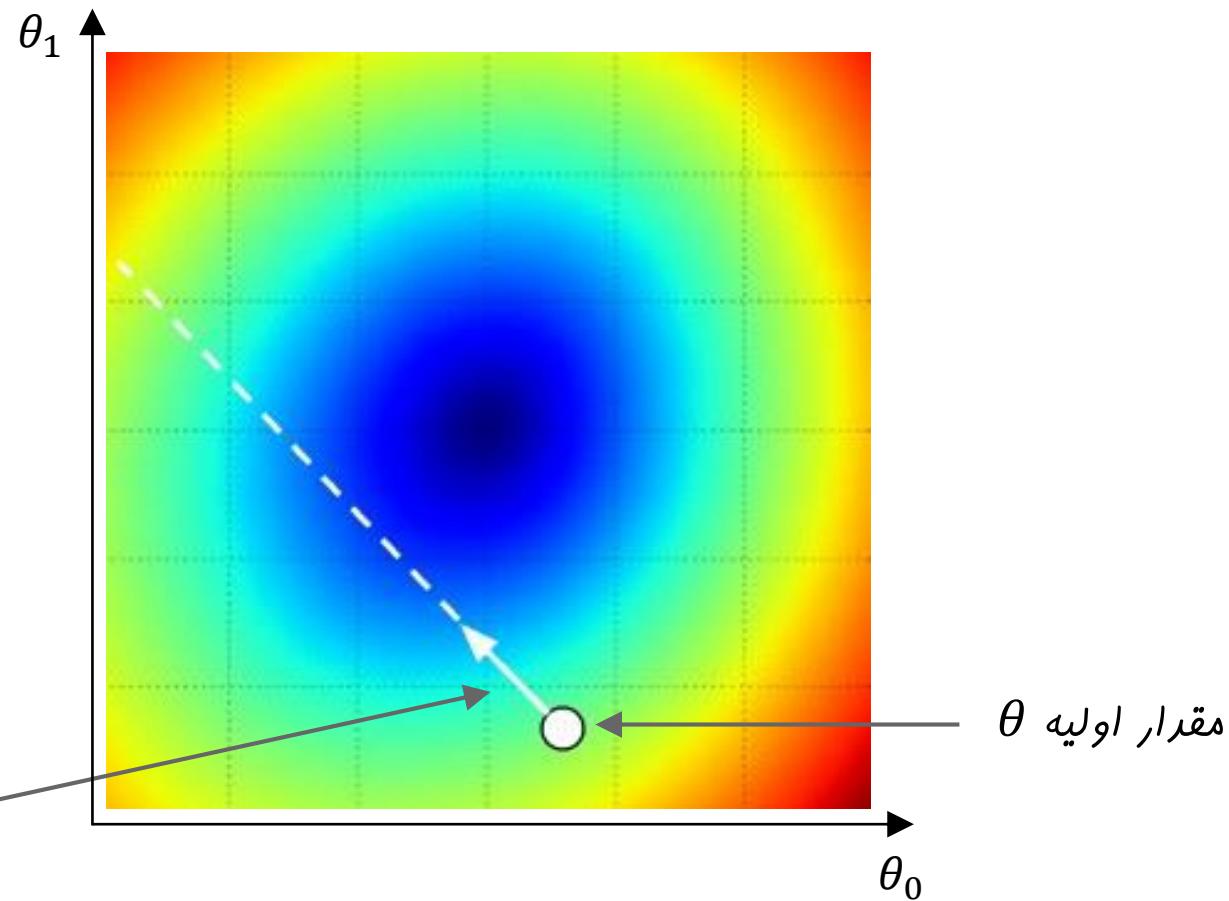
به روز رسانی در هر تکرار با استفاده از یک دسته کوچک تصادفی از نمونه‌ها

يادآوری: الگوریتم گرادیان کاہشی

۵۳

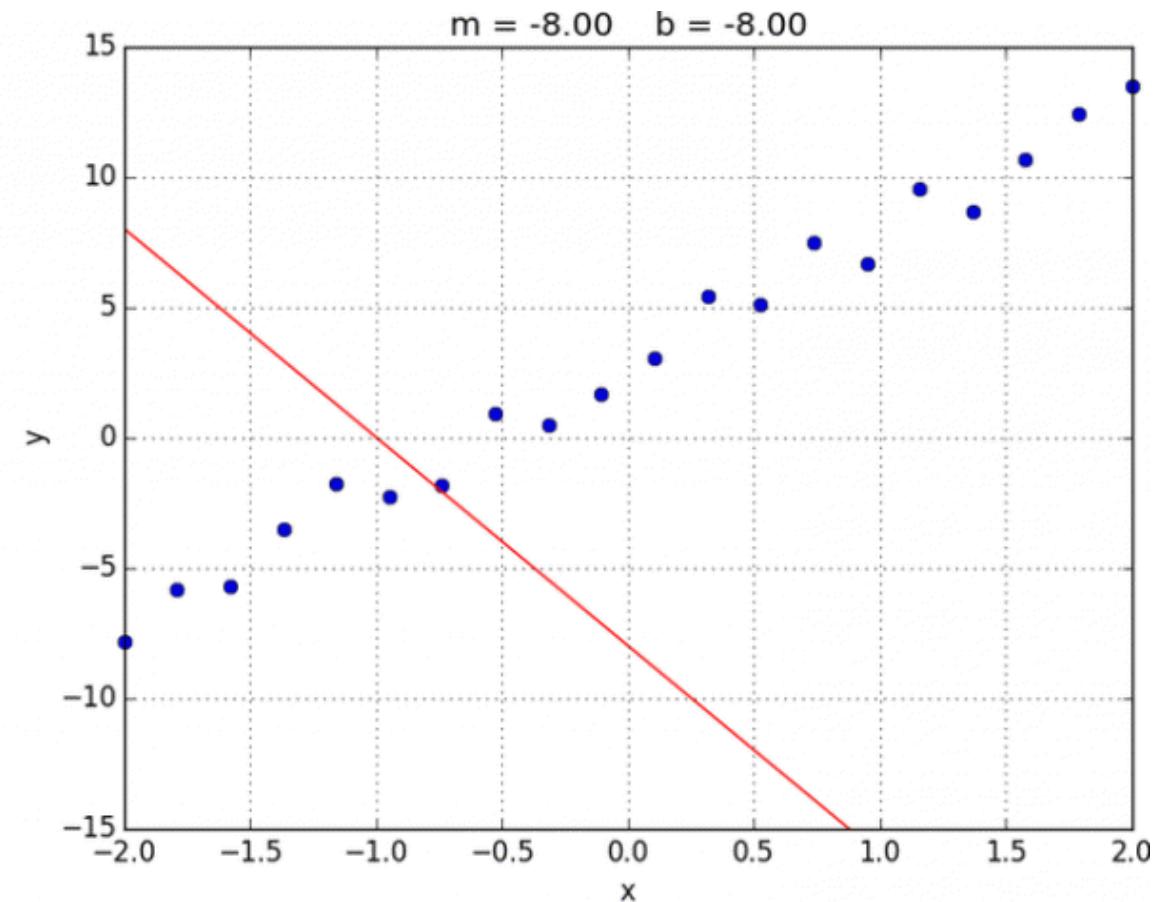
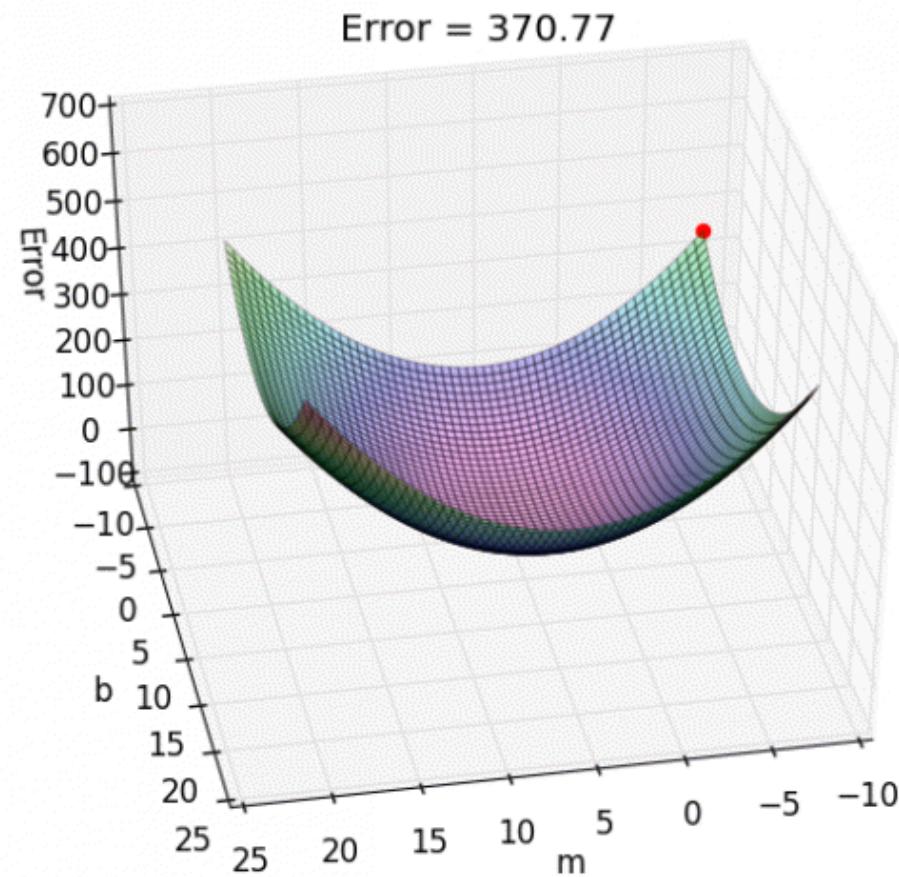
$$\theta_j = \theta_j - \alpha \left(\frac{\partial J}{\partial \theta_j} \right)$$

فلاف بجهت گرادیان



يادآوری: الگوریتم گرادیان کاہشی

۵۴



ڪرسيون خطي پند مڌيڙه

(گرسیون تک متغیره (یک ویژگی)

۵۶

□ رگرسیون خطی با یک ویژگی.

متراز (فوت مربع) x	قیمت (۱۰۰۰ دلار) y
۲۱۰۴	۴۶۰
۱۴۱۶	۲۳۲
۱۵۳۴	۳۱۵
۸۵۲	۱۷۸
...	...

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

اگریوشن پنڈ متغیره (چند ویژگی)

۵۷

□ رگرسیون خطی با چند ویژگی.

متراز (فوت مربع) x_1	تعداد خوابها x_2	تعداد طبقات x_3	سن خانه (سال) x_4	قیمت (۱۰۰۰ دلار) y
۲۱۰۴	۵	۱	۴۵	۴۶۰
۱۴۱۶	۳	۲	۴۰	۲۳۲
۱۵۳۴	۳	۲	۳۰	۳۱۵
۸۵۲	۲	۱	۳۶	۱۷۸
...

$$h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

اگرسیون پنده متغیره (پنده ویژگی)

۵۸

قیمت (۱۰۰۰ دلار) y	سن خانه (سال) x_4	تعداد طبقات x_3	تعداد اتاق خواب ها x_2	متراژ (فوت مربع) x_1
۴۶۰	۴۵	۱	۵	۲۱۰۴
۲۳۲	۴۰	۲	۳	۱۴۱۶
۳۱۵	۳۰	۲	۳	۱۵۳۴
۱۷۸	۳۶	۱	۲	۸۵۲
...

تعداد نمونه های آموزشی

m

□ نمادها.

$$x^{(2)} = \begin{bmatrix} 1416 \\ 3 \\ 2 \\ 40 \end{bmatrix} \quad \begin{array}{l} \leftarrow x_1^{(2)} \\ \leftarrow x_3^{(2)} \end{array}$$

تعداد ویژگی ها
ورودی ها در i امین نمونه آموزشی
مقدار ویژگی زام در i امین نمونه آموزشی

n □

$x^{(i)}$ □

$x_j^{(i)}$ □

فرضیه

۵۹

□ رگرسیون خطی تک متغیره.

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

□ رگرسیون خطی چند متغیره.

$$h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

□ برای سادگی، تعریف می‌کنیم $x_0 = 1$

$$x = \begin{bmatrix} x_0 = 1 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \quad \Rightarrow \quad h_{\theta}(x) = \theta^T x$$

کرادیان کاھشی در گرسیون خطي پند مڌيده

گرادیان کاہشی

۶۱

فرضیه. □

$$h_{\theta}(x) = \theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

پارامترها. □

$$\theta = (\theta_0, \theta_1, \dots, \theta_n)$$

$$J(\theta_0, \theta_1, \dots, \theta_n) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

تابع هزینه. □

repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

}

$$\frac{\partial}{\partial \theta_j} J(\theta) = \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}$$

ترفندهای عملی در اکسیون پند متخیله

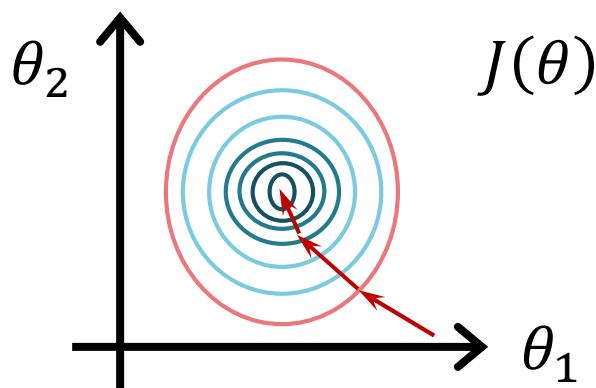
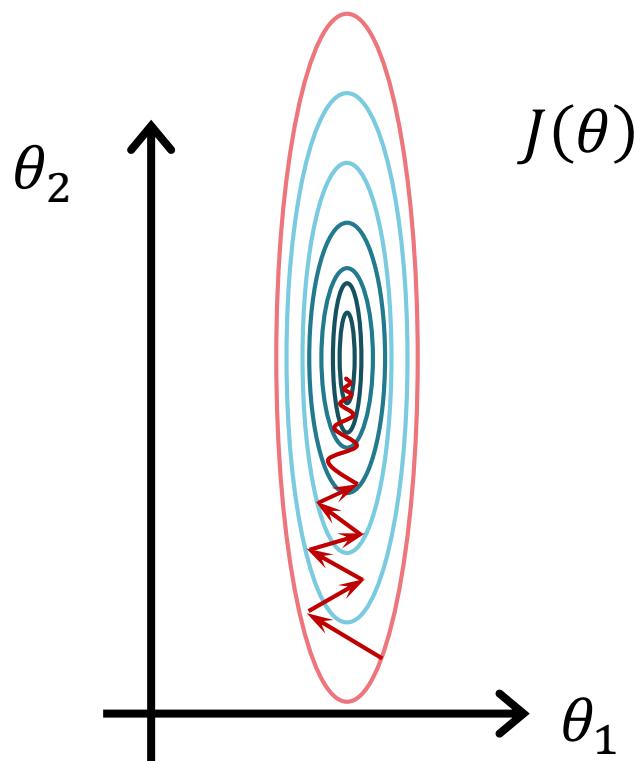
مقیاس‌بندی ویژگی‌ها

تعیین نرخ یادگیری

مقیاس‌بندی ویژگی‌ها (نرمال‌سازی)

۶۳

- ایده. اطمینان از این که مقادیر ویژگی‌ها در یک مقیاس مشابه قرار دارند.
- هدف. افزایش سرعت همگرایی در گرادیان کاهشی.
- مثال.



مقیاس‌بندی ویژگی‌ها

۶۴

□ مقیاس‌بندی. مقدار هر ویژگی در ضریب کوچکی از بازه $[-1, 1]$ قرار دارد.

$$x_1 = \frac{\text{size} - 1000}{2000} \quad -0.5 \leq x_1 \leq 0.5 \quad \text{مثال.} \quad \square$$

$$x_2 = \frac{\# \text{bedrooms} - 2}{5} \quad -0.5 \leq x_2 \leq 0.5$$

□ نرمال‌سازی میانگین.

$$x_j = \frac{x_j - \mu_j}{\sigma_j}$$

میانگین \rightarrow
انحراف معیار \rightarrow

گرادیان کاہشی

۶۵

□ گرادیان کاہشی.

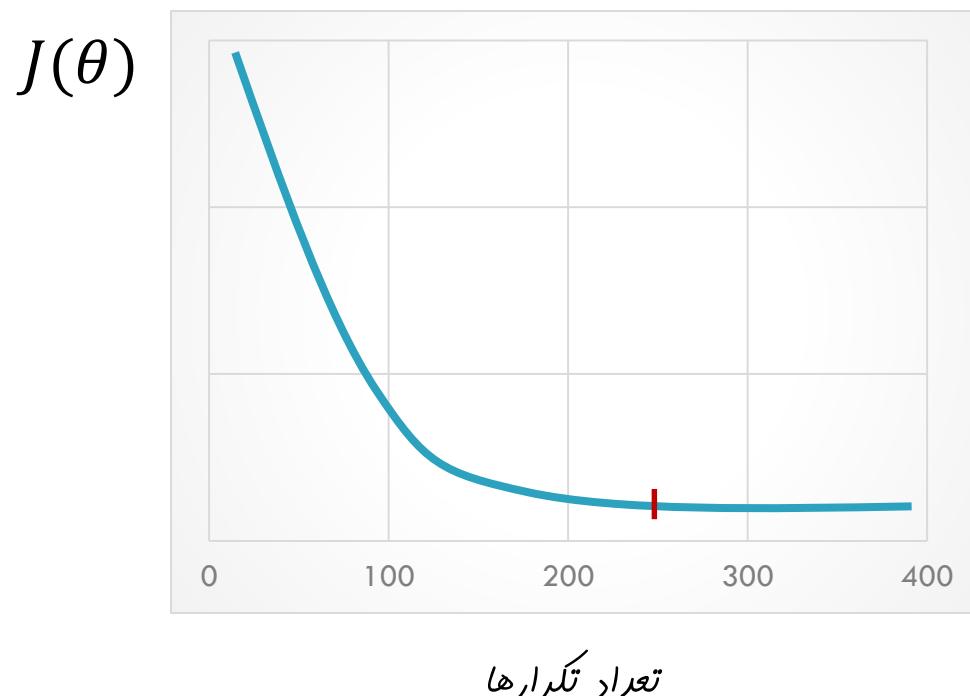
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

- س. چگونه می‌توان مطمئن شد گرادیان کاہشی به درستی عمل می‌کند؟
- س. مقدار مناسب برای نرخ یادگیری چیست؟

عملکرد صمیح برای گرادیان کاوشی

۶۶

- آزمایش همگرایی.
- اگر مقدار $J(\theta)$ در یک تکرار به اندازه‌ای کمتر از 10^{-3} تغییر کند، همگرایی رخ داده است.

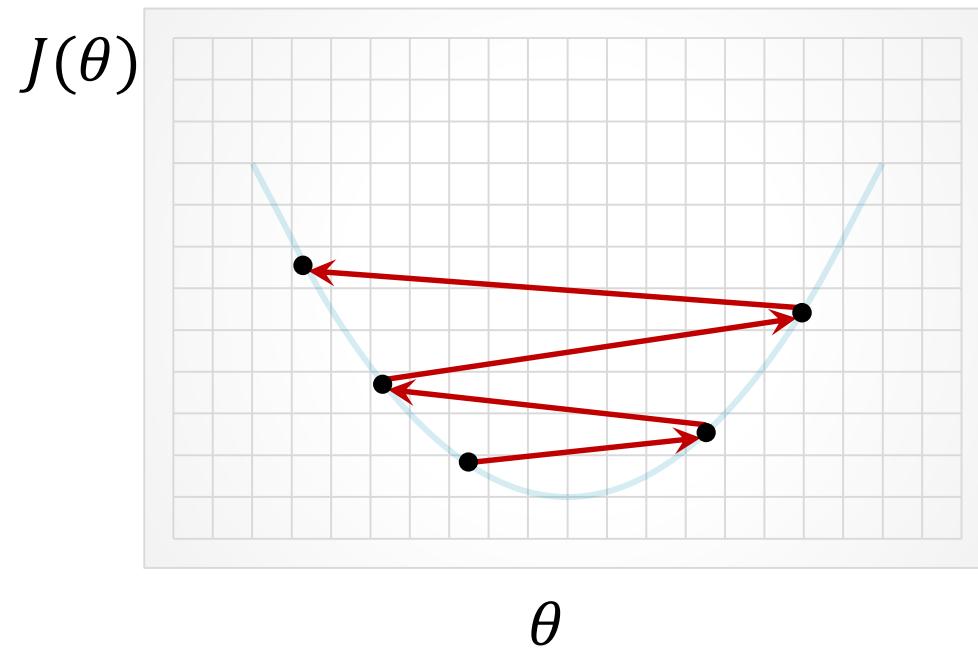


نیشانه‌های عملکرد نادرست گرادیان کاهشی

۶۷

□ عدم همگرایی.

□ راه حل. از مقادیر کوچک‌تری برای نرخ یادگیری استفاده کن، اما اگر نرخ یادگیری بیش از حد کوچک باشد همگرایی بسیار کند خواهد بود.



خلاصه

۶۸

- نرخ یادگیری.
- بسیار کوچک: همگرایی بسیار کند
- بسیار بزرگ: همگرایی کند یا عدم همگرایی
- انتخاب نرخ یادگیری.
- به منظور انتخاب یک مقدار مناسب برای نرخ یادگیری، مقادیر زیر را امتحان کنید:
..., 0.001, 0.003, 0.01, 0.03, .01, 0.3, 1.0, ...

رسیون پندارلای

قیمت‌گذاری یک خانه: انتفاب ویژگی‌ها

۷.

$$h_{\theta}(x) = \theta_0 + \theta_1 \times (\text{frontage}) + \theta_2 \times (\text{depth})$$

x_1 x_2

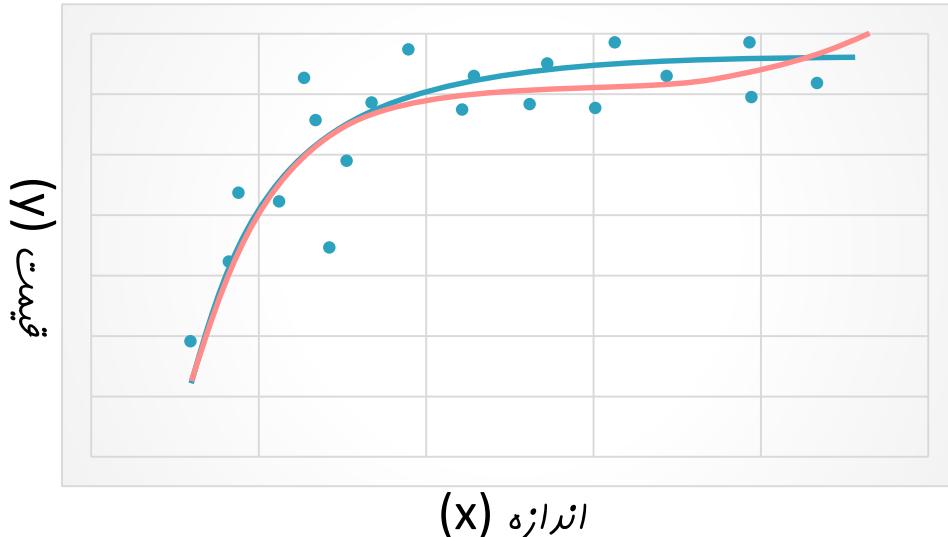
$$\text{Area} = \text{frontage} \times \text{depth}$$



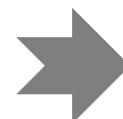
$$h_{\theta}(x) = \theta_0 + \theta_1 \times (\text{Area})$$

رگرسیون پندهای

۷۱



$$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$



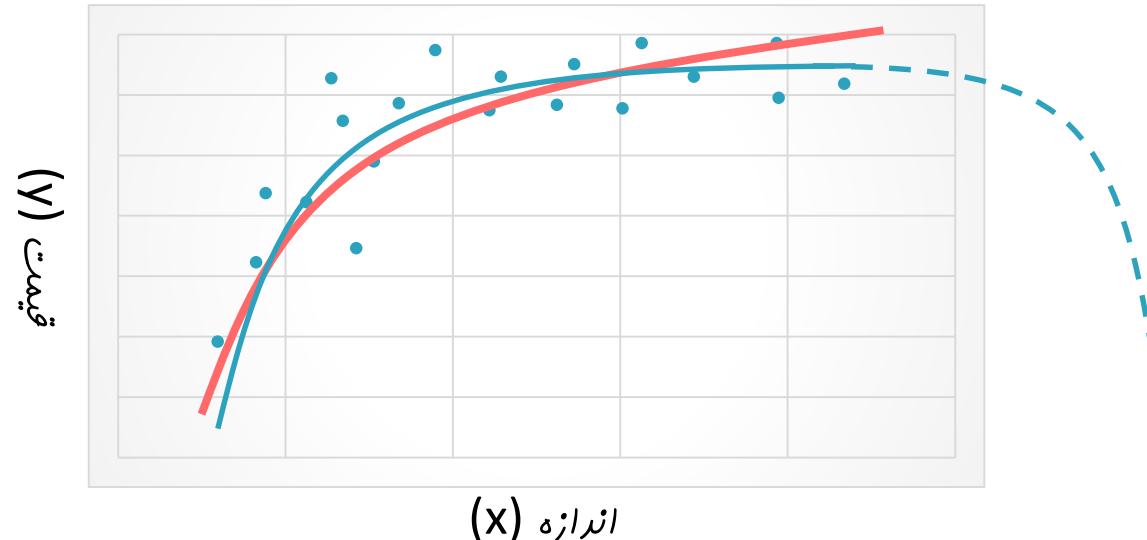
$$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

$$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

$$\begin{aligned} x_1 &= (\text{Size})^1 & 1 \leq x_1 \leq 10^3 \\ x_2 &= (\text{Size})^2 & 1 \leq x_2 \leq 10^6 \\ x_3 &= (\text{Size})^3 & 1 \leq x_3 \leq 10^9 \end{aligned}$$

انتفاب ویژگی‌ها

۷۲



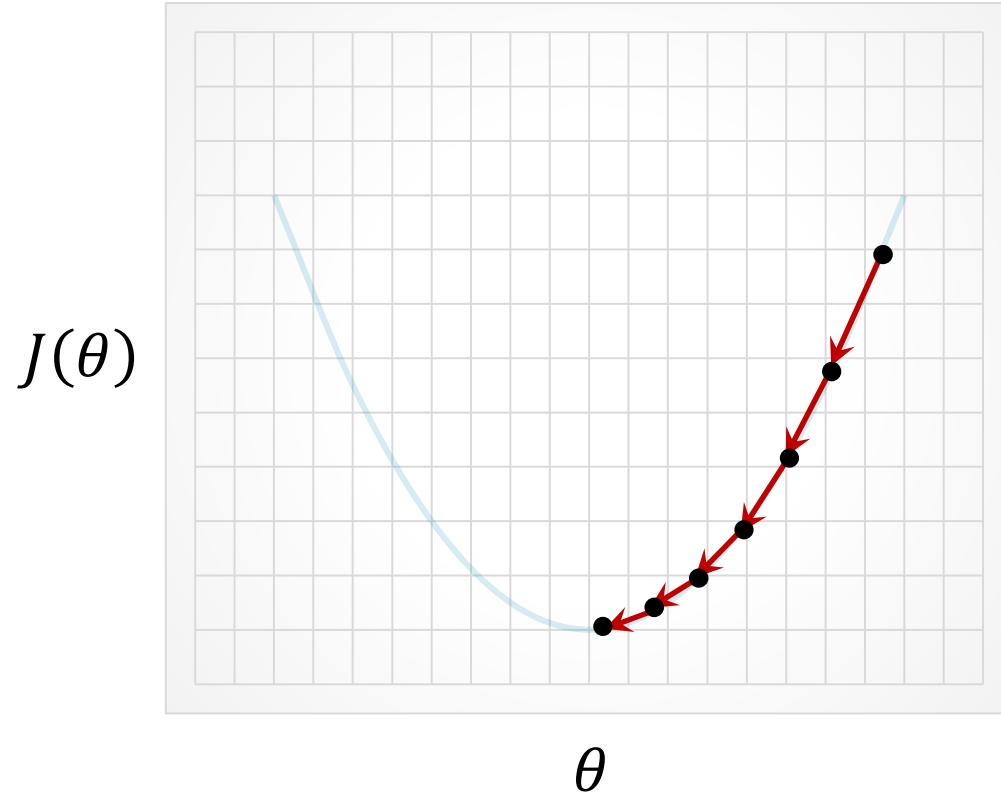
$$h_{\theta}(x) = \theta_0 + \theta_1(size) + \theta_2(size)^2$$

$$h_{\theta}(x) = \theta_0 + \theta_1(size) + \theta_2(\sqrt{size})$$

رسیون خطي: معادله نرمال

گرادیان کاهشی و معادله نرمال

۷۴



□ معادله نرمال. یک روش تحلیلی به منظور تعیین مقدار پارامترها.

□ گرادیان کاهشی.

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

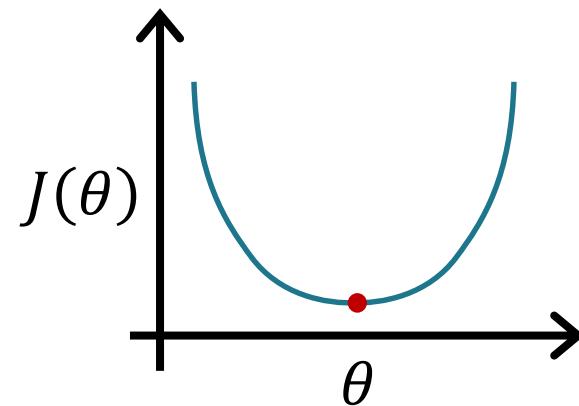
قاعده به روز رسانی

معادله نرمال

۷۵

$$\theta \in \mathbb{R}: J(\theta) = a\theta^2 + b\theta + c$$

$$\frac{d}{d\theta} J(\theta) \stackrel{\text{def}}{=} 0$$



$$\theta \in \mathbb{R}^{n+1}: J(\theta_0, \theta_1, \dots, \theta_n) = \frac{1}{2} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

$$\frac{\partial}{\partial \theta_j} J(\theta) \stackrel{\text{def}}{=} 0 \quad (j = 0, 1, 2, \dots, n)$$

معلمات نرمال: مثال (m = 4)

٧٦

x_0	Size (feet ²) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price (\$1000) y
1	2104	5	1	45	460
1	1416	3	2	40	232
1	1534	3	2	30	315
1	852	2	1	36	178

$$X = \begin{bmatrix} 1 & 2104 & 5 & 1 & 45 \\ 1 & 1416 & 3 & 2 & 40 \\ 1 & 1534 & 3 & 2 & 30 \\ 1 & 852 & 2 & 1 & 36 \end{bmatrix}$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix} \quad y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \end{bmatrix}$$

$$X\theta = y$$

معلمات نرمال: مثال (m = 5)

٧٧

x_0	Size (feet ²) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price (\$1000) y
1	2104	5	1	45	460
1	1416	3	2	40	232
1	1534	3	2	30	315
1	852	2	1	36	178
1	3000	4	1	38	540

$$X = \begin{bmatrix} 1 & 2104 & 5 & 1 & 45 \\ 1 & 1416 & 3 & 2 & 40 \\ 1 & 1534 & 3 & 2 & 30 \\ 1 & 852 & 2 & 1 & 36 \\ 1 & 3000 & 4 & 1 & 38 \end{bmatrix}$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix} \quad y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \\ 540 \end{bmatrix}$$

$$X\theta = y$$

معادله نرمال: مالت کلی

۷۸

$$X\theta = y$$

m نمونه آموزشی؛ n ویژگی \square

$$x^{(i)} = \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix} \in \mathbb{R}^{n+1} \quad X = \begin{bmatrix} \cdots (x^{(1)})^T \cdots \\ \cdots (x^{(2)})^T \cdots \\ \cdots (x^{(3)})^T \cdots \\ \vdots \\ \cdots (x^{(m)})^T \cdots \end{bmatrix} \in \mathbb{R}^{m \times (n+1)} \quad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ \vdots \\ y^{(m)} \end{bmatrix} \in \mathbb{R}^m$$

ماتریس طراحی

معادله نرمال

۷۹

□ حل دستگاه معادلات خطی.

$$X\theta = y$$

$$X^T X \theta = X^T y \quad \leftarrow \text{معادله نرمال}$$

$$\theta = \underbrace{(X^T X)^{-1} X^T}_{X^+} y$$

Python:

```
theta = pinv(X.T @ X) @ X.T @ y
```

گرادیان کاھشی و معادله نرمال

معادله نرمال

گرادیان کاھشی

- عدم نیاز به انتخاب α
- عدم نیاز به تکرار

- نیاز به انتخاب α
- نیاز به تکرارهای زیاد

$X^T X$ به دلیل نیاز به محاسبه معکوس ماتریس X حتی برای مقادیر بسیار بزرگ n به خوبی کار می‌کند.

$$n < 10000$$

$$n \geq 10000$$

معادله نرمال و معکوس ناپذیر

۸۲

□ معادله نرمال.

$$\theta = (X^T X)^{-1} X^T y$$

□ س. اما اگر $X^T X$ معکوس پذیر نباشد چه می شود؟

Python: `theta = pinv(X.T @ X) @ X.T @ y`

↑
شبه معکوس

علل معمکوس ناپذیری

۸۳

□ افزونگی ویژگی‌ها. [وابستگی خطی]

$$x_1 = \text{size}(\text{feet}^2)$$

$$x_2 = \text{size}(m^2)$$

$$x_1 = (3.28)^2 x_2$$

□ تعداد بسیار زیاد ویژگی‌ها. [$n \geq m$]

□ راه حل. حذف برخی از ویژگی‌ها با استفاده از تنظیم [در ادامه]

رسیون با وزن دهنده محلی

(و)ش‌های یادگیری پارامتری و غیرپارامتری

۸۵

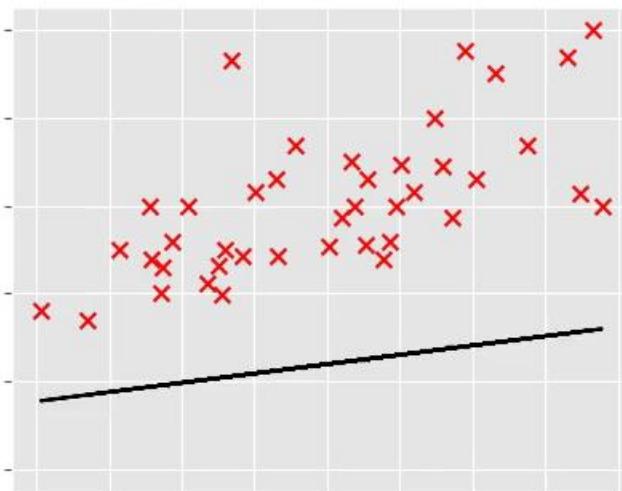
□ روش‌های یادگیری پارامتری.

□ یک مجموعه ثابت از پارامترها وجود دارد.

□ برای پیش‌بینی داده‌های جدید، نیازی به مجموعه آموزشی نداریم.

□ مثال: رگرسیون، و رگرسیون لجستیک، شبکه‌های عصبی.

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}$$



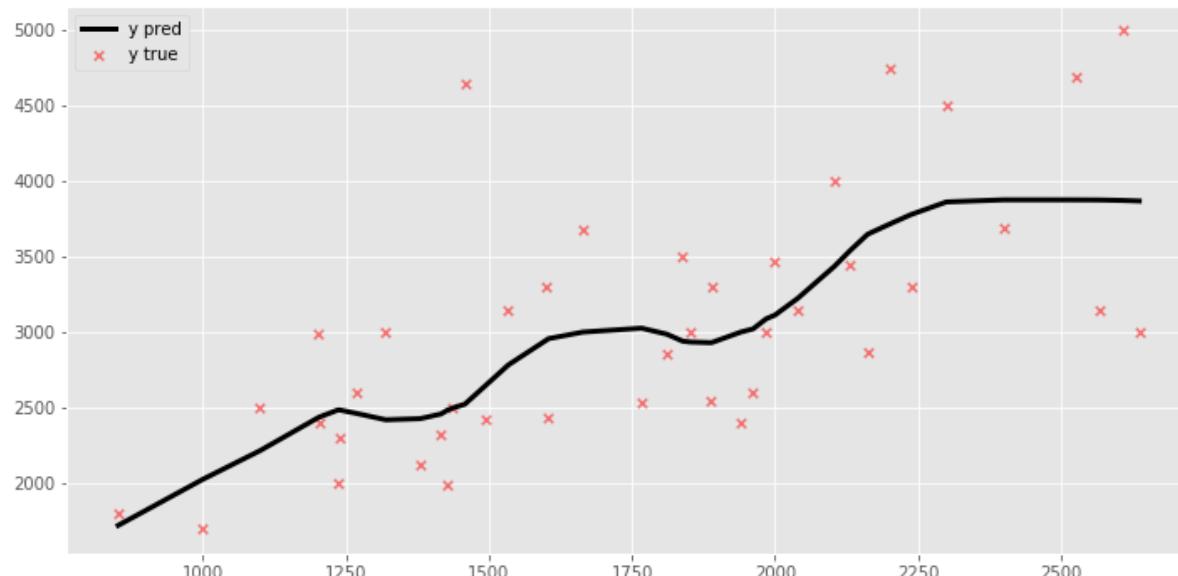
$$h_{\theta}(x^{(new)}) = \theta^T x^{(new)}$$

(و)ش‌های یادگیری پارامتری و غیرپارامتری

۸۶

روش‌های یادگیری غیرپارامتری.

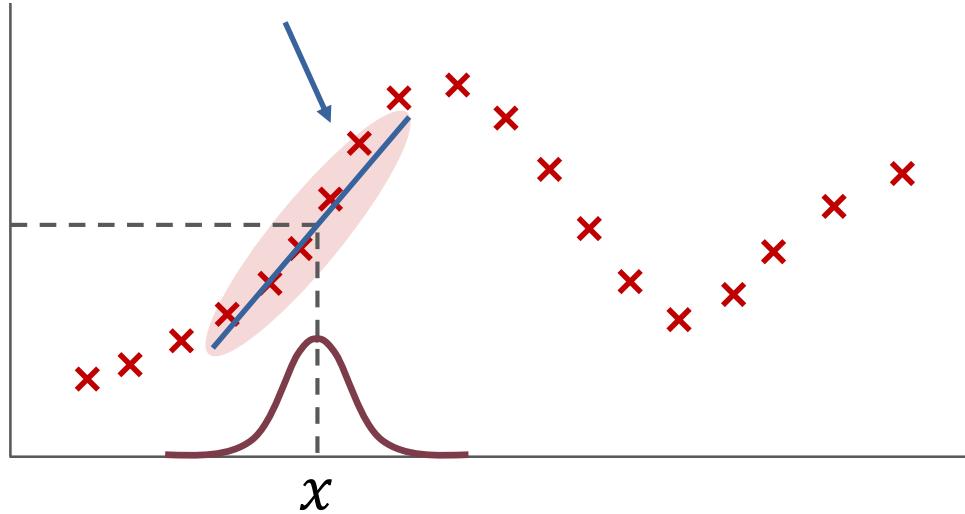
- تعداد پارامترها با افزایش اندازه مجموعه آموزشی (به صورت خطی) افزایش می‌یابد.
- به منظور پیش‌بینی داده‌های جدید، به تمام مجموعه آموزشی نیاز داریم.
- مثال: رگرسیون با وزن‌دهی محلی [صفحه بعد](#).



اگرسيون با وزن دهی محلی

۸۷

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$



ایده. دادن اهمیت بیشتر به داده‌های نزدیک‌تر. □

$$w^{(i)} = \exp \left(-\frac{(x^{(i)} - x)^2}{2\tau^2} \right)$$

پهنای باند

$$J(\theta) = \sum_{i=1}^m w^{(i)} \left(y^{(i)} - h_{\theta}(x^{(i)}) \right)^2$$

رگرسیون: تفسیر احتمالاتی

(گرسیون: تفسیر احتمالاتی

۸۹

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

خطا

مدل.

خطا.

$$\epsilon^{(i)} \sim N(0, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

- در نظر گرفتن تأثیر عوامل دیگر
- مانند ویژگی‌های درنظر گرفته نشده
- در نظر گرفتن تأثیر نویز.

□ خطاهای مستقل هستند و از یک توزیع یکسان گاووسی پیروی می‌کنند. [iid]

$$y^{(i)} \sim N(\theta^T x^{(i)}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

تَفْهِمِين بِيُشْتَرِين درستنمايی

۹۰

□ تابع درستنمايی. احتمال مشاهده داده‌های آموزشی به عنوان تابعی از پارامترهای θ

$$L(\theta) = p(Y|X; \theta) = \prod_{i=1}^m \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$\begin{aligned} l(\theta) &= \ln L(\theta) = \ln \prod_{i=1}^m \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) && \text{□ لگاريتم تابع درستنمايی.} \\ &= \sum_{i=1}^m \ln \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \right] \\ &= m \ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2} \sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2 \end{aligned}$$

تخمین بیشترین درست‌نمایی

۹۱

□ تخمین بیشترین درست‌نمایی.

□ انتخاب یک مقدار برای پارامتر θ به گونه‌ای که $l(\theta)$ بیشینه گردد.

$$\begin{aligned}\theta_{MLE} &= \arg \max_{\theta} l(\theta) \\ &= \arg \max_{\theta} m \ln \frac{1}{\sigma \sqrt{2\pi}} - \frac{1}{2\sigma^2} \sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2 \\ &= \arg \max_{\theta} -\frac{1}{2} \sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2 \\ &= \arg \min_{\theta} \frac{1}{2} \sum_{i=1}^m (y^{(i)} - \theta^T x^{(i)})^2 \quad \xleftarrow{\text{تابع هزینه}}\end{aligned}$$

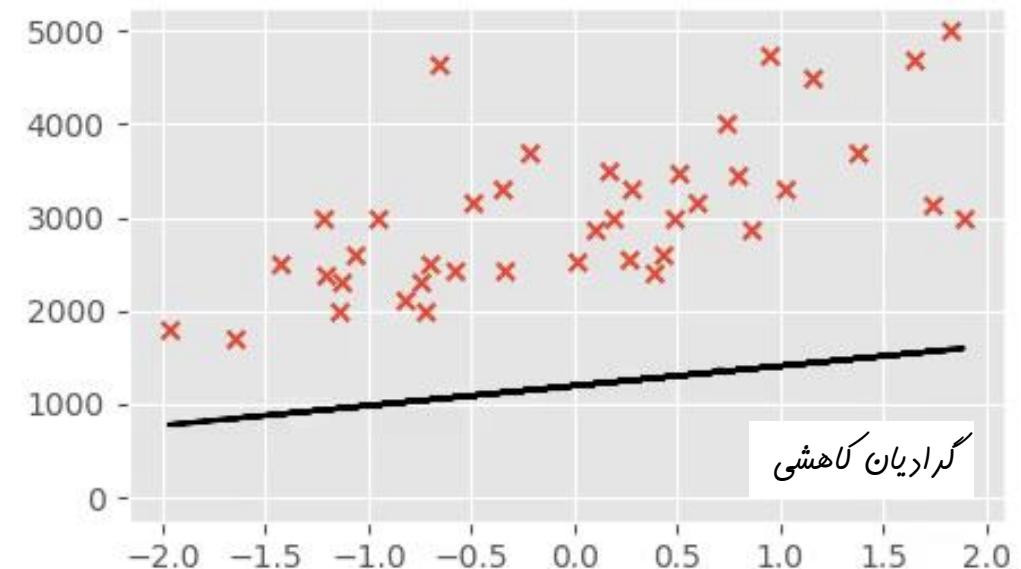
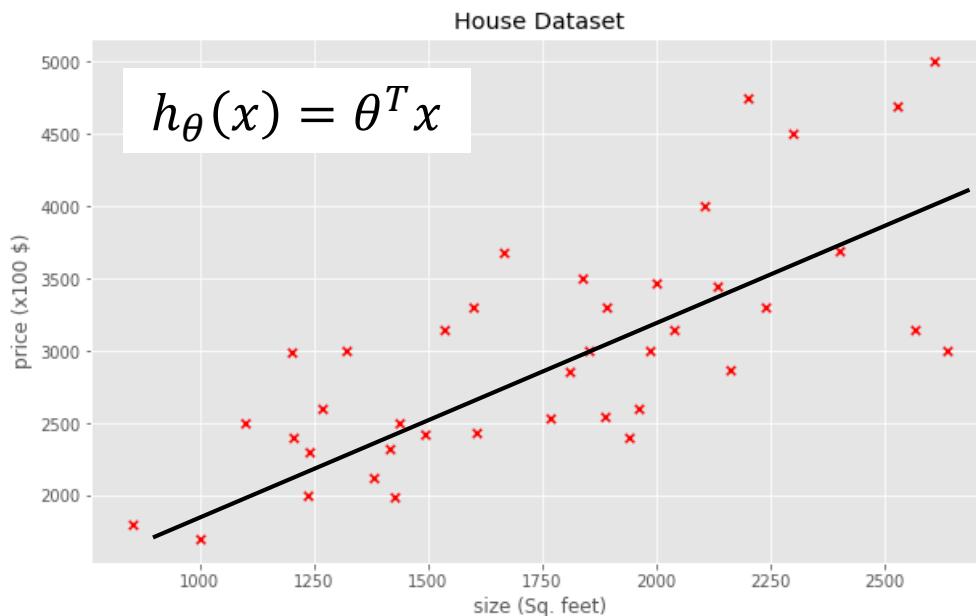
دسته‌بندی: اگرنسیون لجیستیک

سید ناصر رضوی n.razavi@tabrizu.ac.ir

یادآوری: گرسیون

۲

هدف. تخمین یک کمیت پیوسته با توجه به مقادیر ویژگی‌ها. □



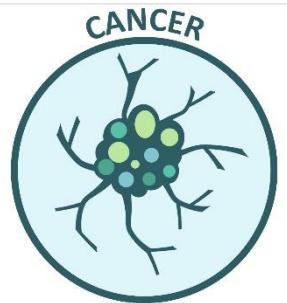
تابع هزینه

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

کمینه‌سازی

دسته‌بندی

۳



ایمیل: هرزنامه (بله / خیر?)

تراکنش برخط: کلاهبرداری (بله / خیر?)

غده سرطانی: خوشخیم / بدخیم?

در این مثال‌ها، متغیری که می‌خواهیم مقدارش را پیش‌بینی کنیم دارای دو مقدار است:

$$y \in \{0,1\}$$

صفر: «کلاس منفی» (مانند غده خوش‌خیم)

یک: «کلاس مثبت» (مانند غده بدخیم)

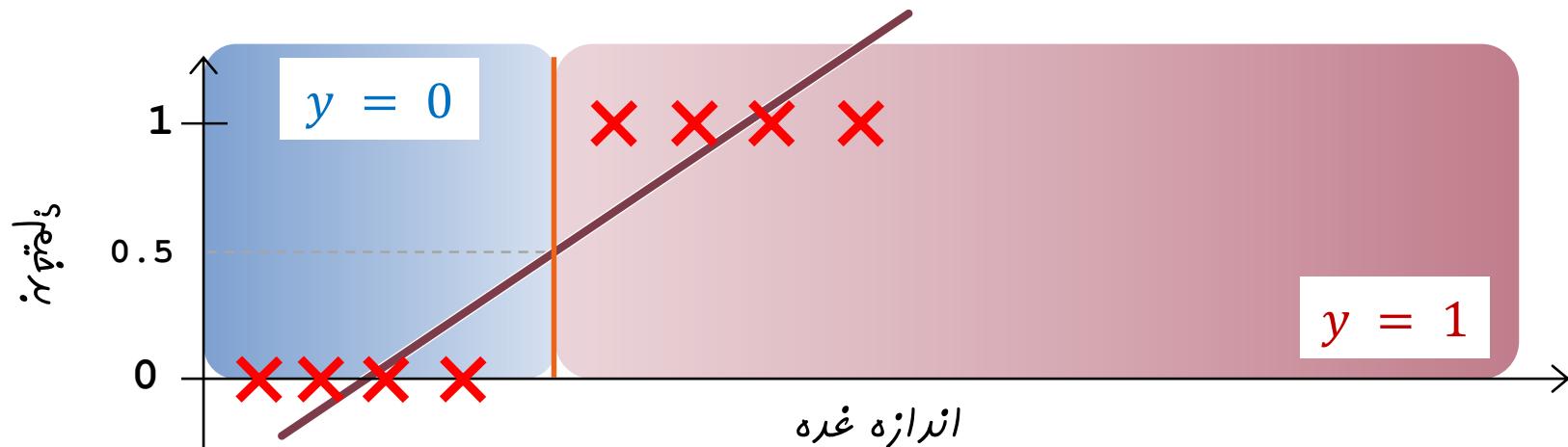
دسته‌بندی. پیش‌بینی یک متغیر با مقادیر گستته.

دسته‌بندی دودویی

دسته‌بندی چندکلاسی

دسته‌بندی

۴



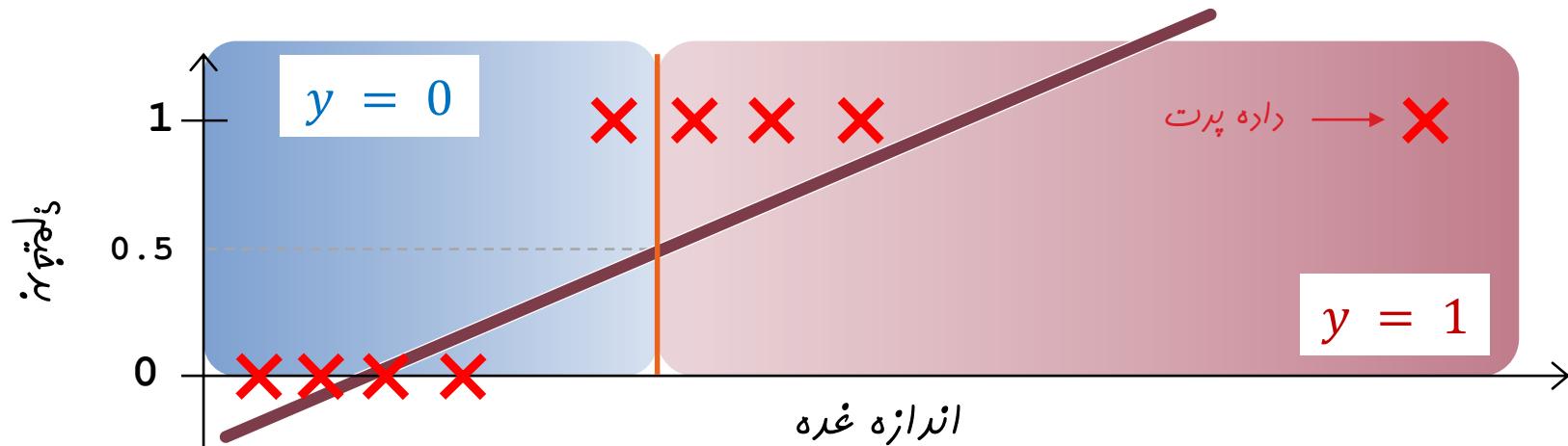
□ قرار دادن یک آستانه بر روی خروجی دسته‌بند:

□ اگر $y = 1$, آنگاه $h_\theta(x) \geq 0.5$

□ اگر $y = 0$, آنگاه $h_\theta(x) < 0.5$

دسته‌بندی

۵



□ قرار دادن یک آستانه بر روی خروجی دسته‌بند:

□ اگر $y = 1$, آنگاه $h_\theta(x) \geq 0.5$

□ اگر $y = 0$, آنگاه $h_\theta(x) < 0.5$

دسته‌بندی

۶

در دسته‌بندی دودویی داریم:

$$y = 0 \text{ یا } y = 1$$

اما در رگرسیون ممکن است:

$$h_{\theta}(x) < 0 \text{ یا } h_{\theta}(x) > 1$$

رگرسیون لجستیکی (دسته‌بندی).

$$0 \leq h_{\theta}(x) \leq 1$$

بازنمایی فرضیه در (گرسیون لمستیک

بازنمایی فرضیه

^

هدف. □

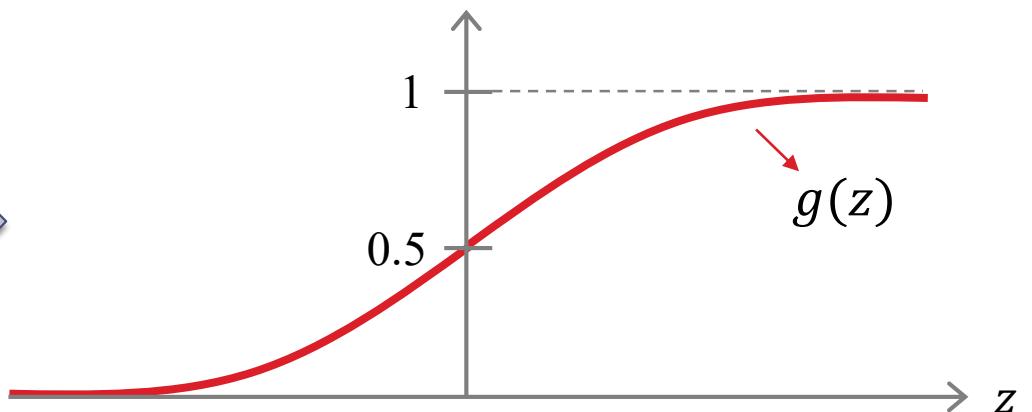
$$0 \leq h_{\theta}(x) \leq 1$$

فرضیه. □

$$h_{\theta}(x) = g(\theta^T x)$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

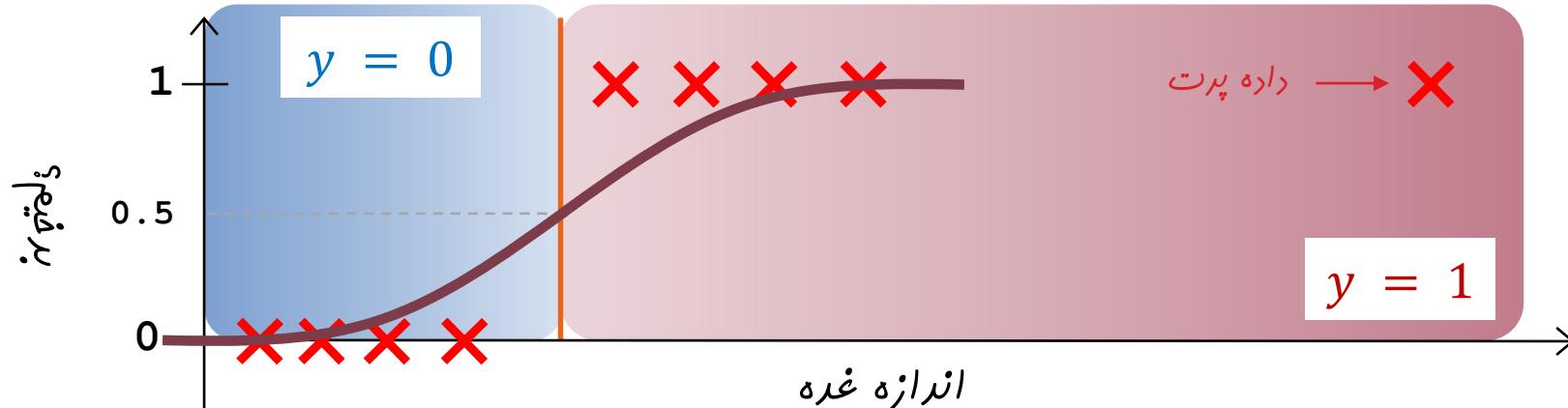
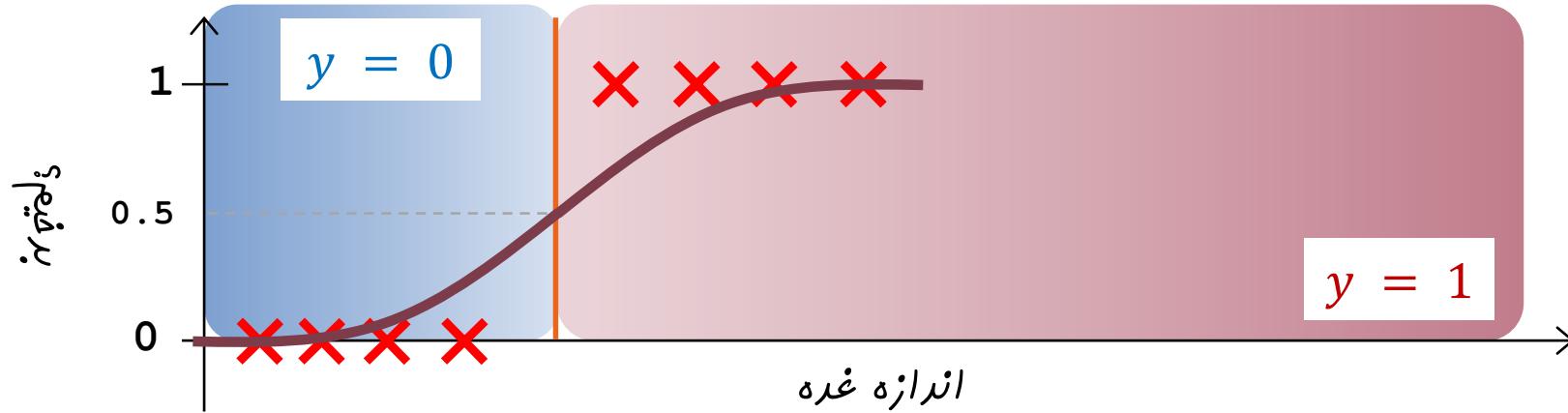
تابع سیگموید
(لجستیک)



$$0 \leq g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \leq 1$$

اگریوں لجستیک و دسائیوندی

۹



فرضیه

۱۰

□ تفسیر خروجی فرضیه.

«احتمال این که ورودی x به دسته $y = 1$ تعلق داشته باشد»

$x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ tumorSize \end{bmatrix}, \quad h_{\theta}(x) = 0.7$ □ مثال. اگر داشته باشیم:

در این صورت، به احتمال ۷۰ درصد، این غده سرطانی بدخیم است.

$$p(y = 1|x; \theta) = h_{\theta}(x)$$

$$p(y = 0|x; \theta) = 1 - p(y = 1|x; \theta) = 1 - h_{\theta}(x)$$

فرضیه

۱۱

$$p(y = 1|x; \theta) = h_{\theta}(x)$$

$$p(y = 0|x; \theta) = 1 - h_{\theta}(x)$$

$$p(y|x; \theta) = h_{\theta}(x)^y (1 - h_{\theta}(x))^{1-y}$$

$$\begin{aligned} L(\theta) &= p(Y|X; \theta) = \prod_{i=1}^m p(y^{(i)}|x^{(i)}; \theta) \\ &= \prod_{i=1}^m h_{\theta}(x^{(i)})^{y^{(i)}} (1 - h_{\theta}(x^{(i)}))^{1-y^{(i)}} \end{aligned}$$

□ تفسیر احتمالاتی فرضیه.

□ تابع درستنمایی.

تَفْهِمِين بِيُشْتَرِين درستنمايی

۱۲

□ لگاریتم تابع درستنمايی.

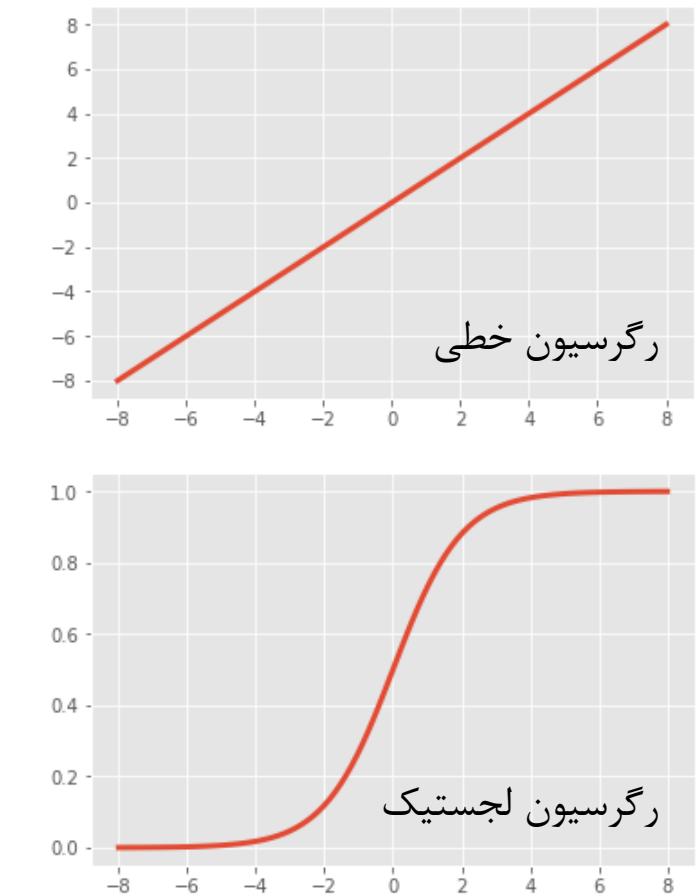
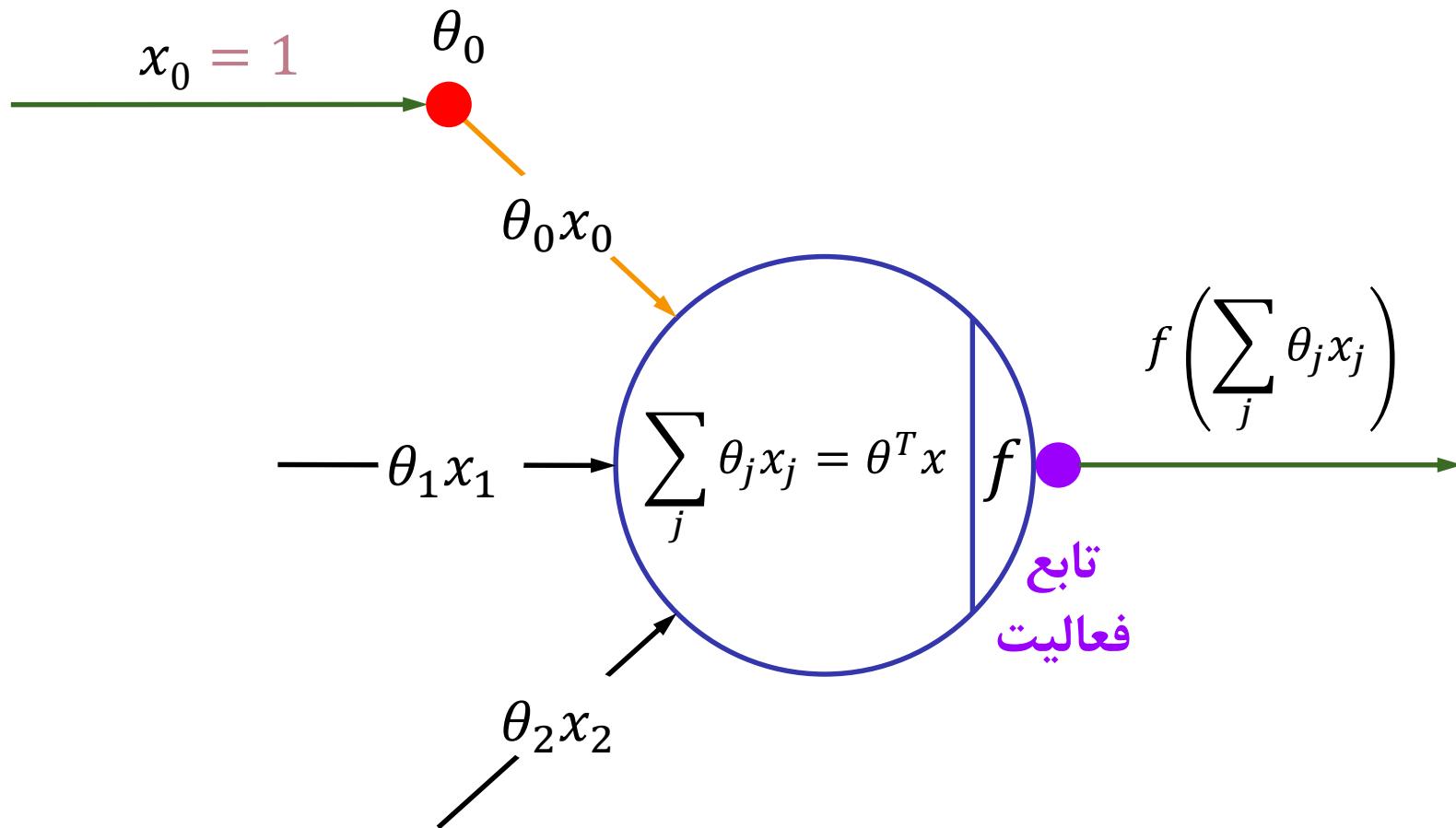
$$\begin{aligned} l(\theta) &= \log L(\theta) = \log \prod_{i=1}^m h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}} \\ &= \sum_{i=1}^m \log \left(h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}} \right) \\ &= \sum_{i=1}^m y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)})) \end{aligned}$$

□ تابع هزینه.

$$J(\theta) = -l(\theta) = \sum_{i=1}^m -y^{(i)} \log h_\theta(x^{(i)}) - (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))$$

رگرسیون خطی و رگرسیون لجستیک

۱۲



درز تصدیم‌گیری

مرز تضمیم‌گیری

۱۵

□ رگرسیون لجستیک.

$$h_{\theta}(x) = g(\theta^T x)$$

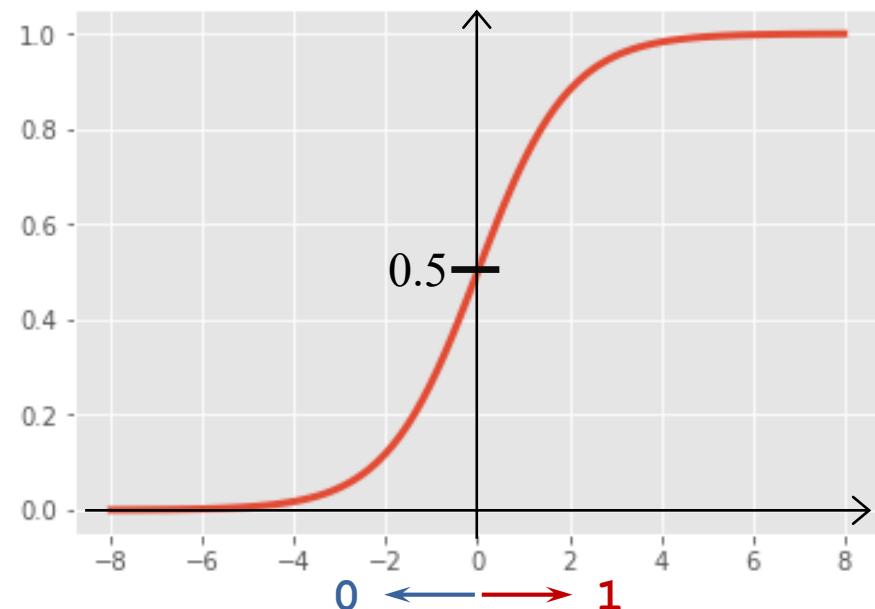
□ قرار دادن یک آستانه بر روی خروجی دسته‌بند:

$$y = 1: h_{\theta}(x) \geq 0.5 \Rightarrow \theta^T x \geq 0$$

$$y = 0: h_{\theta}(x) < 0.5 \Rightarrow \theta^T x < 0$$

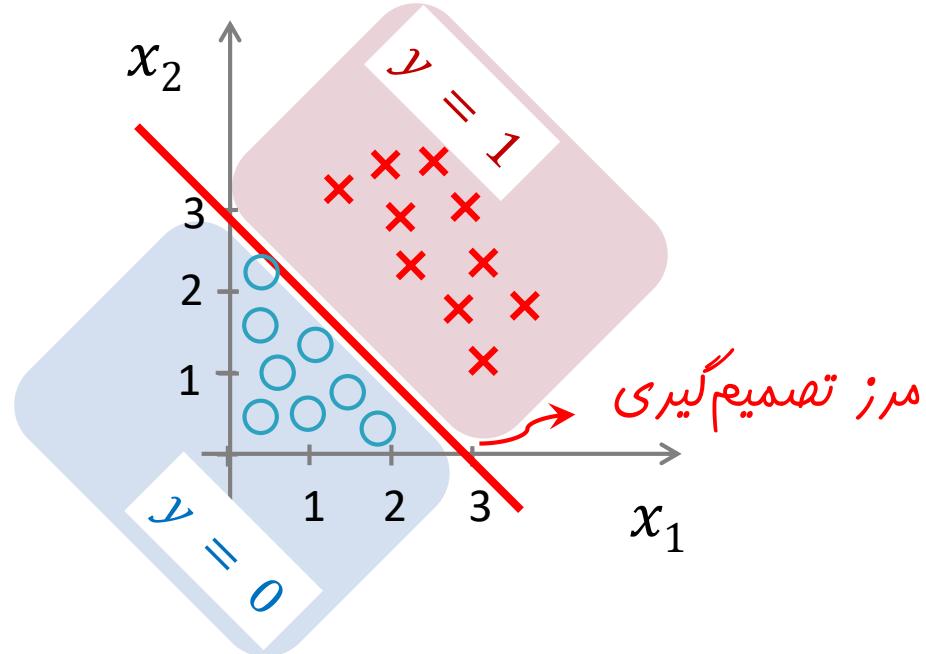
$$\theta^T x = 0$$

معادله مرز تضمیم‌گیری



مرز تصمیم‌گیری

۱۶



مرز تصمیم‌گیری

□ مرز تصمیم‌گیری.

$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

-3 1 1

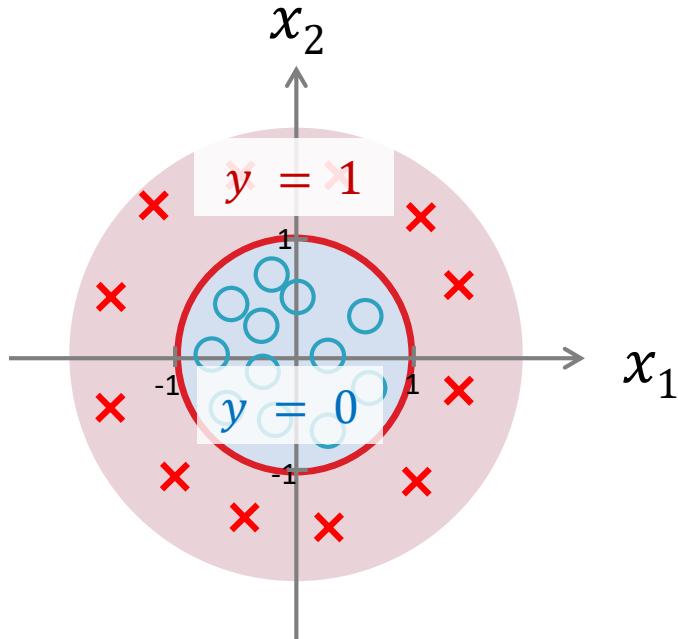
□ خروجی y برابر با ۱ است، اگر $-3 + x_1 + x_2 \geq 0$

□ $x_1 + x_2 \geq 3 \Rightarrow y = 1$

□ $x_1 + x_2 < 3 \Rightarrow y = 0$

درز تصمیم‌گیری غیرخطی

۱۷



مرز تصمیم‌گیری.

$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

↓ ↓ ↓ ↓ ↓
-1 0 0 1 1

$$x_1^2 + x_2^2 \geq 1 \Rightarrow y = 1$$

$$x_1^2 + x_2^2 < 1 \Rightarrow y = 0$$

تابع هزینه

اگریوں لجستیک

۱۹

□ مجموعه آموزشی.

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$

□ نمونه آموزشی.

$$x \in \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \quad x_0 = 1, \quad y \in \{0,1\}$$

□ فرضیه.

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

□ س. مقادیر پارامترهای θ را چگونه انتخاب کنیم؟

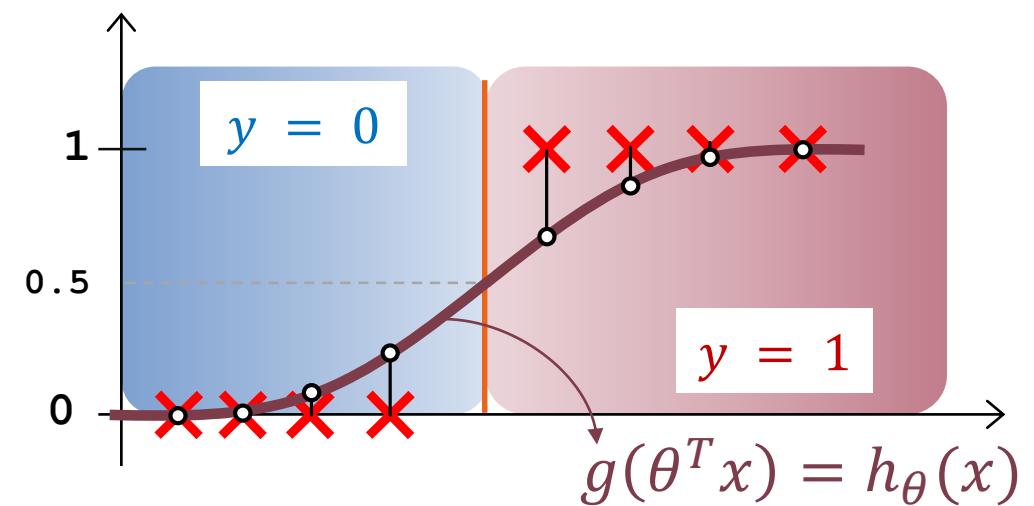
اگرسيون لمسيك

۲۰

□ تابع هزینه.

$$J(\theta) = \sum_{i=1}^m \frac{1}{2} (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

$$cost(h_{\theta}(x^{(i)}), y^{(i)}) = \frac{1}{2} (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

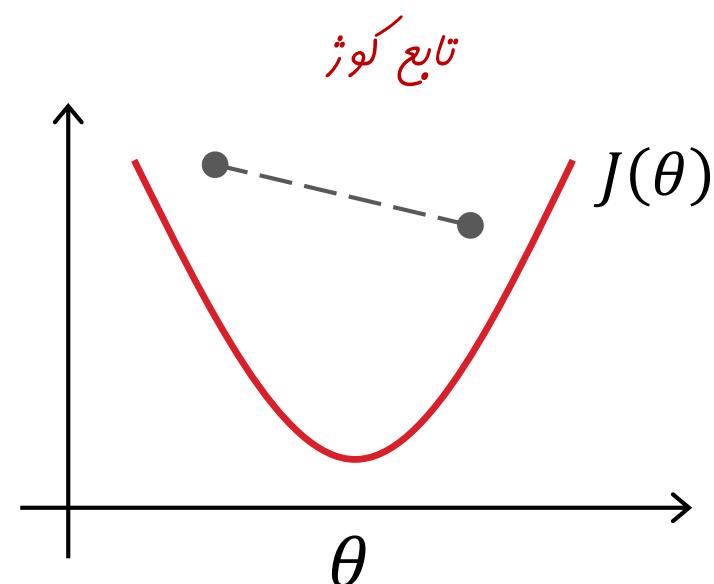
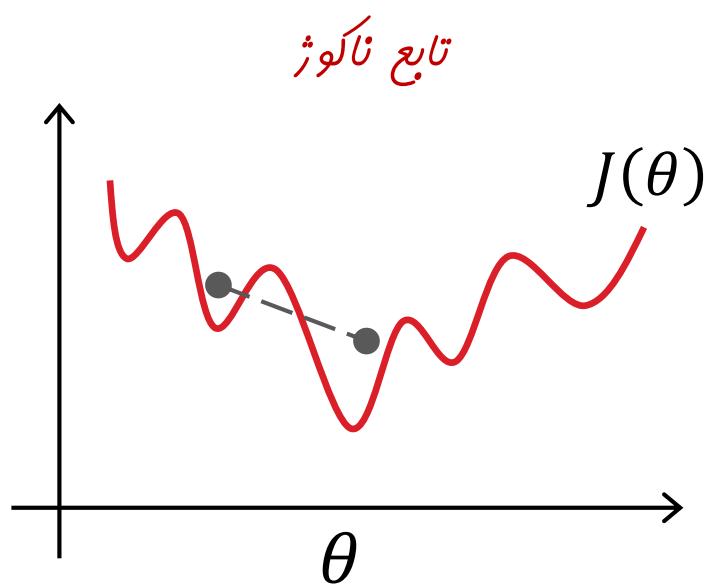


توجه. از آنجا که $h_{\theta}(x^{(i)})$ یک تابع غیرخطی از پارامترها است، تابع هزینه دیگر یک تابع کوز نخواهد بود.

تابع هزینه

۲۱

□ توابع کوثر و ناکوثر.

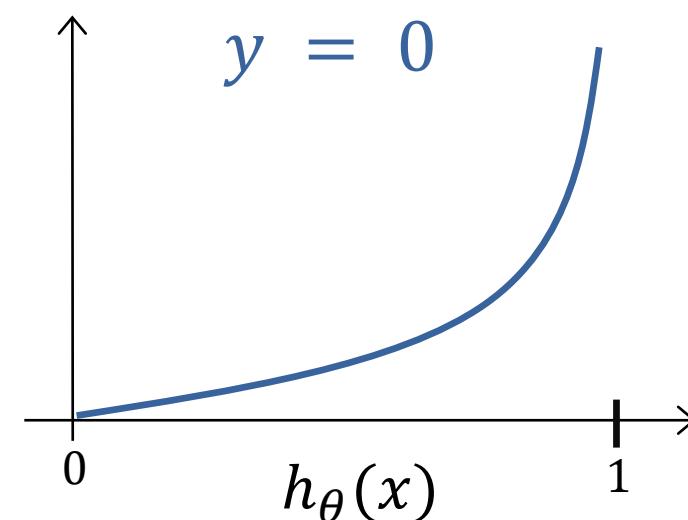
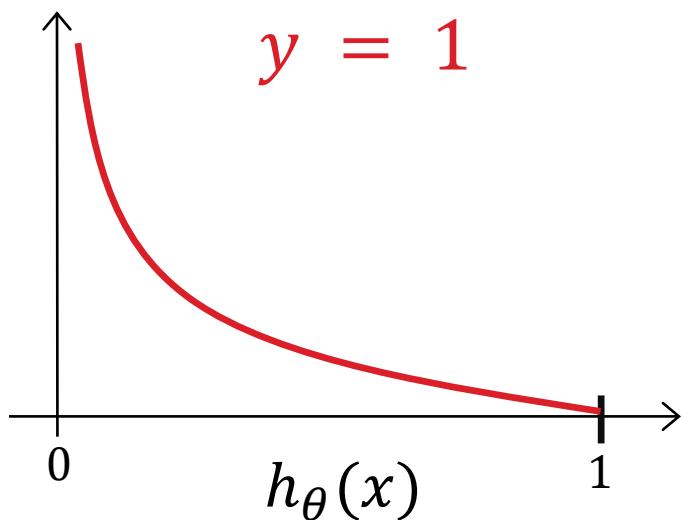


تابع هزینه در آگریشن لجستیک

۲۲

□ تابع هزینه.

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)), & y = 1 \\ -\log(1 - h_\theta(x)), & y = 0 \end{cases}$$



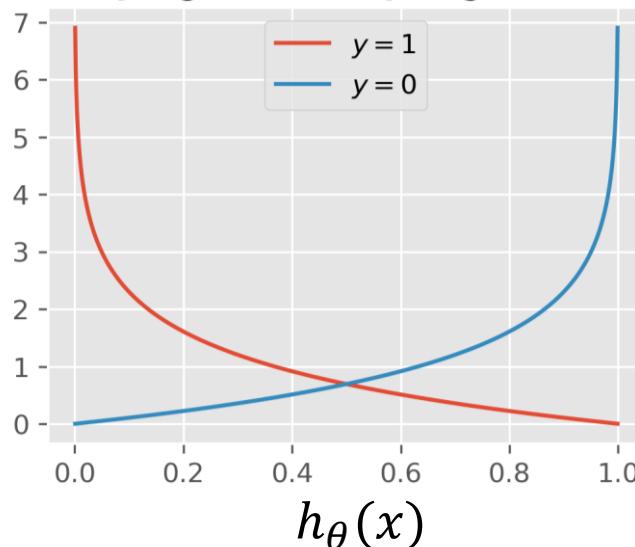
تابع هزینه در آگریشن لجستیک

۲۳

ساده‌سازی تابع هزینه. □

$$J(\theta) = \sum_{i=1}^m cost(h_\theta(x^{(i)}), y^{(i)})$$

$$cost(h_\theta(x^{(i)}), y^{(i)}) = -y^{(i)} \log(h_\theta(x^{(i)})) - (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))$$



تابع هزینه در آگریشن لجستیک

۲۴

□ تابع هزینه.

$$\begin{aligned} J(\theta) &= \sum_{i=1}^m \text{cost}(h_\theta(x^{(i)}), y^{(i)}) \\ &= \sum_{i=1}^m \left[-y^{(i)} \log(h_\theta(x^{(i)})) - (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right] \end{aligned}$$

$$\min_{\theta} J(\theta)$$

□ تعیین مقدار پارامترها.

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

□ پیش‌بینی برای ورودی جدید x

تابع هزینه در آگریشن لجستیک

۲۵

□ تابع هزینه.

$$J(\theta) = \sum_{i=1}^m \left[-y^{(i)} \log(h_\theta(x^{(i)})) - (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right]$$

$$\nabla J(\theta) = X^T(h_\theta(X) - y) \quad \nabla J(\theta) \in \mathbb{R}^{n+1}$$

$$H = X^T \operatorname{diag}(h_\theta(X)(1 - h_\theta(X))) X \quad H \in \mathbb{R}^{(n+1) \times (n+1)}$$

توجه. ماتریس هسین یک ماتریس **مثبت معین** است، بنابراین تابع هزینه یک **تابع کوز** است.

الگوريتم گراديان کاهشی

۲۶

$$J(\theta) = \sum_{i=1}^m \left[-y^{(i)} \log(h_\theta(x^{(i)})) - (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right]$$

□ الگوريتم گراديان کاهشی. [شكل برداري]

```
repeat until convergence {
```

$$\theta := \theta - \alpha \nabla J(\theta)$$

$$\nabla J(\theta) = X^T(h_\theta(X) - y)$$

الگوريتم گراديان کاهشی

۲۷

$$J(\theta) = \sum_{i=1}^m \left[-y^{(i)} \log(h_\theta(x^{(i)})) - (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right]$$

□ الگوريتم گراديان کاهشی.

repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \quad (j = 0, 1, \dots, n)$$

}



$$\frac{\partial}{\partial \theta_j} J(\theta) = \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}$$

الگوریتم گرادیان کاہشی

۲۸

$$J(\theta) = \sum_{i=1}^m \left[-y^{(i)} \log(h_\theta(x^{(i)})) - (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right]$$

□ الگوریتم گرادیان کاہشی.

```
repeat until convergence {  
     $\theta_j := \theta_j - \alpha \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}$       ( $j = 0, 1, \dots, n$ )  
}
```

توجه. این الگوریتم درست مانند الگوریتم رگرسیون خطی است و تنها تفاوت در تابع فرضیه است.

دسته‌بندی با پنداشتن

دسته‌بندی با پنده دسته

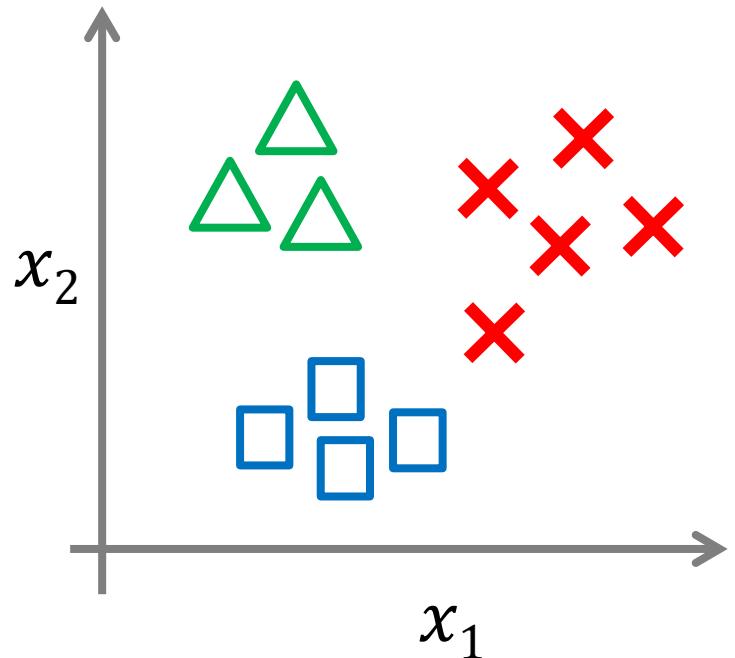
۳۰

□ ایمیل: کاری، خانوادگی، سرگرمی

□ نمودارهای پزشکی: سالم، سرما خوردگی، آنفلوآنزا

□ هوا: آفتابی، ابری، بارانی، برفی

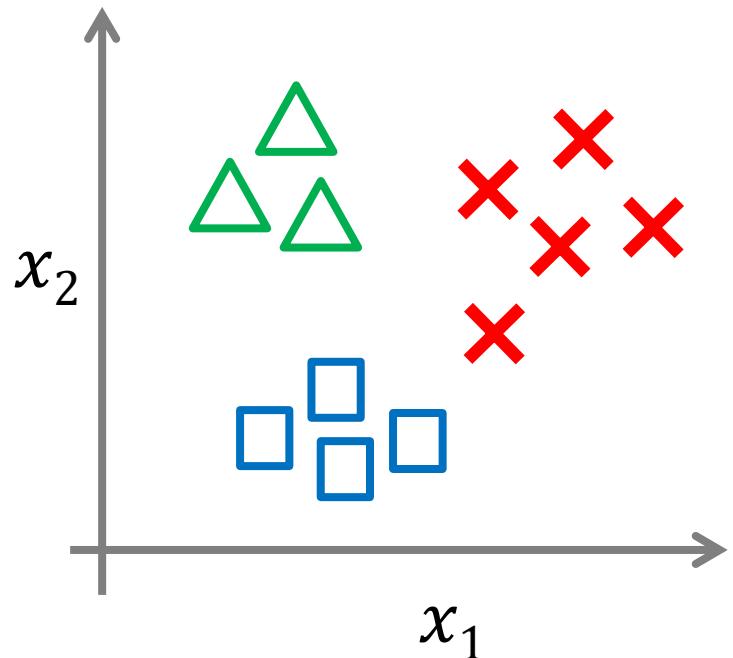
$$y \in \{1, 2, 3, \dots, k\}$$



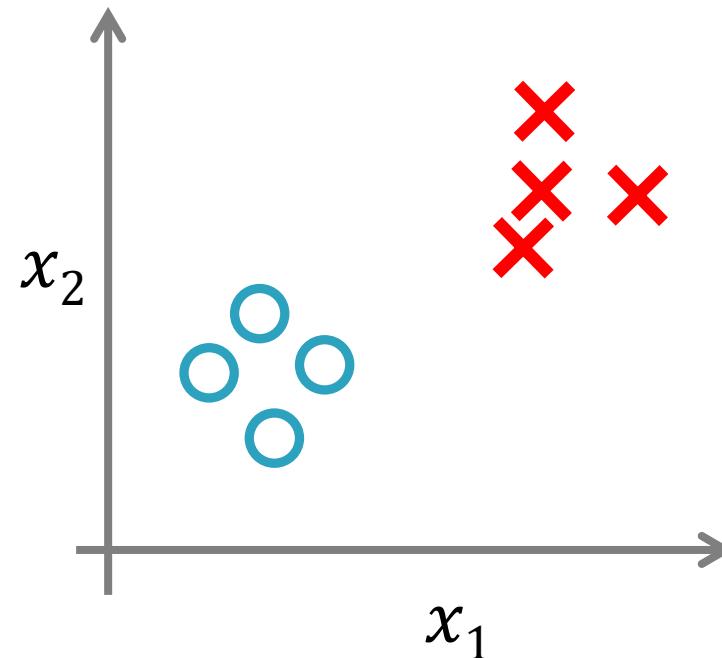
دسته‌بندی با پنداشته

۲۱

دسته‌بندی چند دسته‌ای



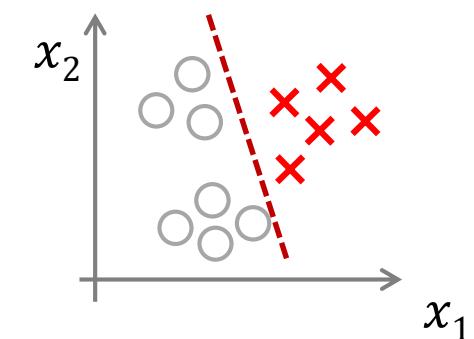
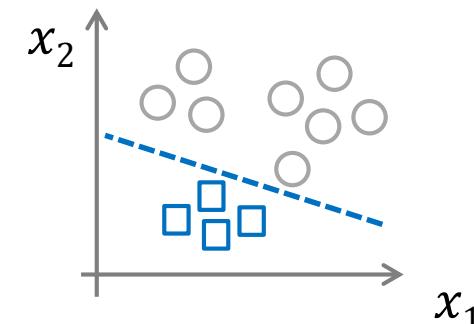
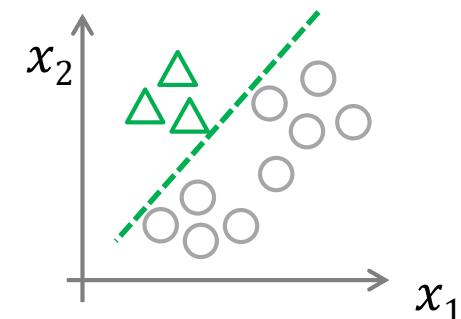
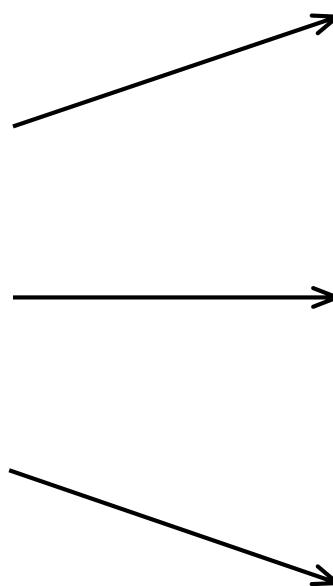
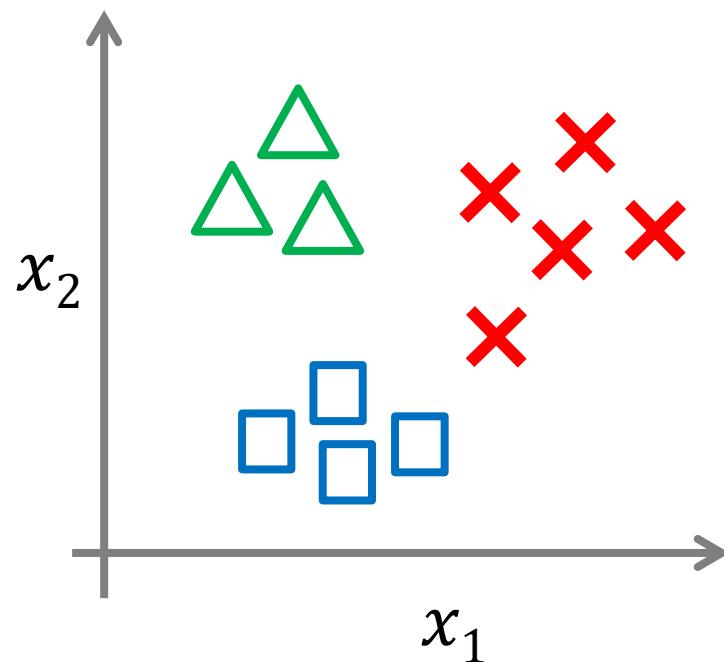
دسته‌بندی دودویی



دسته‌بندی با چند دسته: یکی در برابر همه

۳۲

$$h_{\theta}^{(i)}(x) = P(y = i|x; \theta) \quad (i = 1, 2, 3)$$



- Class 1:
- Class 2:
- Class 3:

دسته‌بندی با پنده دسته: یکی در برابر همه

۳۳

- یکی در برابر همه. به ازای هر دسته i ، دسته‌بند رگرسیون لجیستیک $h_{\theta}^{(i)}(x)$ را به منظور تخمین احتمال تعلق ورودی x به دسته i آموزش بده.
- پیش‌بینی. به منظور دسته‌بندی ورودی جدید x ، دسته i را انتخاب کن به گونه‌ای که:

$$y = \arg \max_i h_{\theta}^{(i)}(x)$$

$$h_{\theta}^{(1)}(x) = 0.25$$

$$h_{\theta}^{(2)}(x) = 0.70$$

$$h_{\theta}^{(3)}(x) = 0.45$$

$$\Rightarrow y = 2$$

روش‌های بهینه‌سازی پیش‌رفته

(وَشْهَايِ بِهِنْهِسَازِيِ پِيشْرِفَتَه

۳۵

□ هدف. یافتن مقدار θ به منظور کمینه‌سازی تابع هزینه.

$$\min_{\theta} J(\theta)$$

□ فرض. برنامه‌ای داریم که با داشتن مقادیر θ , می‌تواند مقادیر زیر را محاسبه کند:

$$J(\theta) \quad \frac{\partial}{\partial \theta_j} J(\theta)$$

repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \quad (j = 0, 1, \dots, n)$$

}

گرادیان کاهشی

(روش‌های بهینه‌سازی پیشرفته)

۳۶

فرض. برنامه‌ای داریم که با داشتن مقادیر θ , می‌تواند مقادیر زیر را محاسبه کند:

$$J(\theta) \quad \frac{\partial}{\partial \theta_j} J(\theta)$$

الگوریتم‌های بهینه‌سازی پیشرفته.

□ گرادیان مزدوج

BFGS □

L-BFGS □

□ مزايا. اين روش‌ها نياز به انتخاب نرخ یادگيري ندارند و عموماً نسبت به الگوریتم گرادیان کاهشي زودتر همگرا می‌شوند.

(وُش‌های بهینه‌سازی پیش‌رفته)

۳۷

مثال. □

$$J(\theta) = (\theta_0 - 5)^2 + (\theta_1 - 5)^2$$



$$\frac{\partial}{\partial \theta_0} J(\theta) = 2(\theta_0 - 5)$$

$$\frac{\partial}{\partial \theta_1} J(\theta) = 2(\theta_1 - 5)$$

```
def J(theta):  
    return (theta[0] - 5) ** 2 + (theta[1] - 5) ** 2
```

```
def grads(theta):  
    return np.array([2 * (theta[0] - 5), 2 * (theta[1] - 5)])
```

(وُش‌های بهینه‌سازی پیش‌رفته)

۳۸

مثال. □

$$J(\theta) = (\theta_0 - 5)^2 + (\theta_1 - 5)^2$$



$$\frac{\partial}{\partial \theta_0} J(\theta) = 2(\theta_0 - 5)$$

$$\frac{\partial}{\partial \theta_1} J(\theta) = 2(\theta_1 - 5)$$

```
from scipy.optimize import minimize
```

```
minimize(J, x0=[0, 0], method='CG', jac=grad)
```

```
fun: 2.477476329894505e-18
jac: array([1.71271335e-08, 1.71271335e-08])
message: 'Optimization terminated successfully.'
nfev: 20
nit: 2
njev: 5
status: 0
success: True
x: array([5., 5.])
```

(وُش‌های بهینه‌سازی پیش‌رفته)

۳۹

مثال. □

$$J(\theta) = (\theta_0 - 5)^2 + (\theta_1 - 5)^2$$



$$\frac{\partial}{\partial \theta_0} J(\theta) = 2(\theta_0 - 5)$$

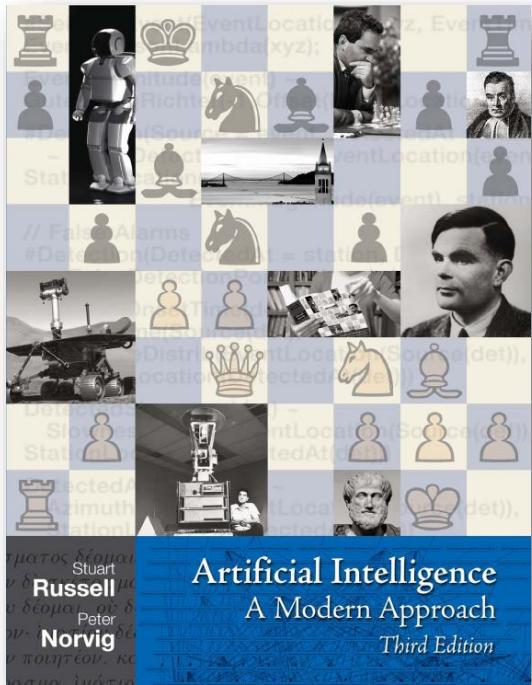
$$\frac{\partial}{\partial \theta_1} J(\theta) = 2(\theta_1 - 5)$$

```
from scipy.optimize import minimize  
  
minimize(J, x0=[0, 0], method='BFGS', jac=
```

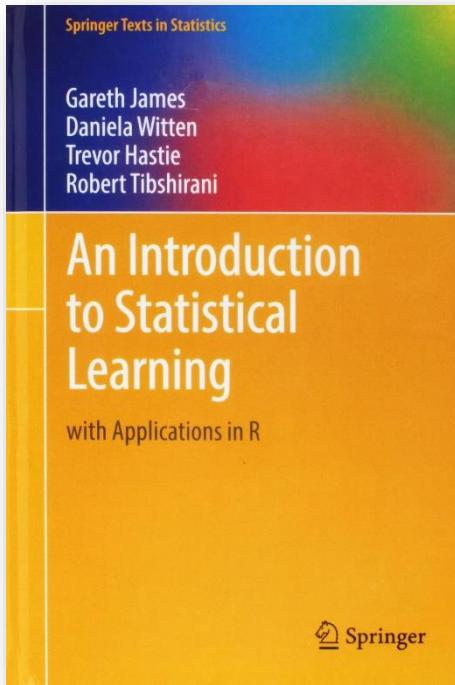
```
fun: 3.5538794606501983e-16  
hess_inv: array([[ 0.75, -0.25],  
                 [-0.25,  0.75]])  
jac: array([-1.17592194e-08, -1.17592194e-08])  
message: 'Optimization terminated successfully.'  
nfev: 16  
nit: 3  
njev: 4  
status: 0  
success: True  
x: array([4.99999999, 4.99999999])
```

مطالعه پیشتر

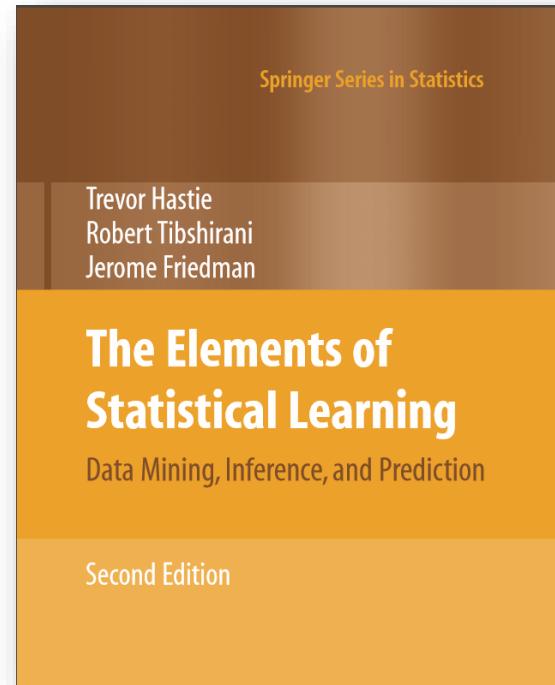
٤٠



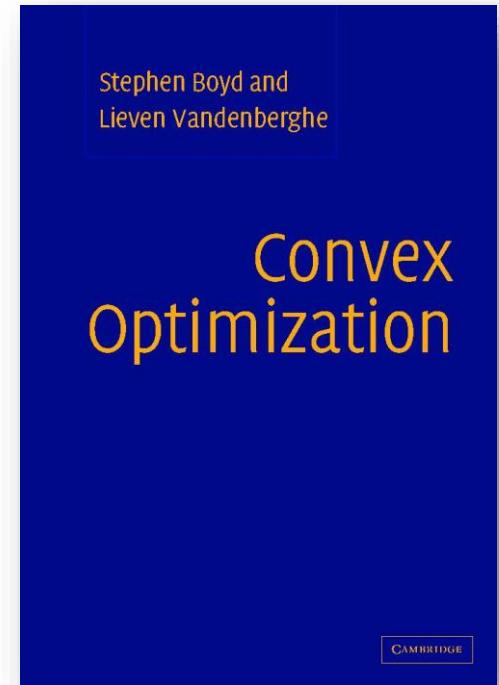
[صفحات ٧٢٧ تا ٧٢٥]



[صفحات ١٣٧ تا ١٣٠]



[صفحات ١٢٨ تا ١١٩]



[بهینه‌سازی محدب]

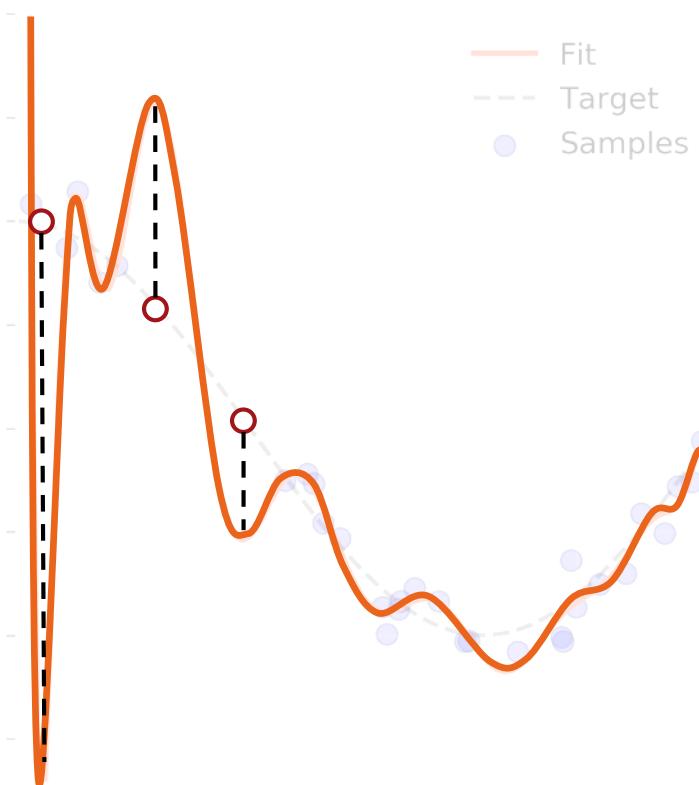
تَنظِيم: بِرْفُورَد بَا بِيَشْ بِرَازِشْ

سید ناصر رضوی www.snrazavi.ir

۱۳۹۷

بیشبرازش و تنظیم

۲

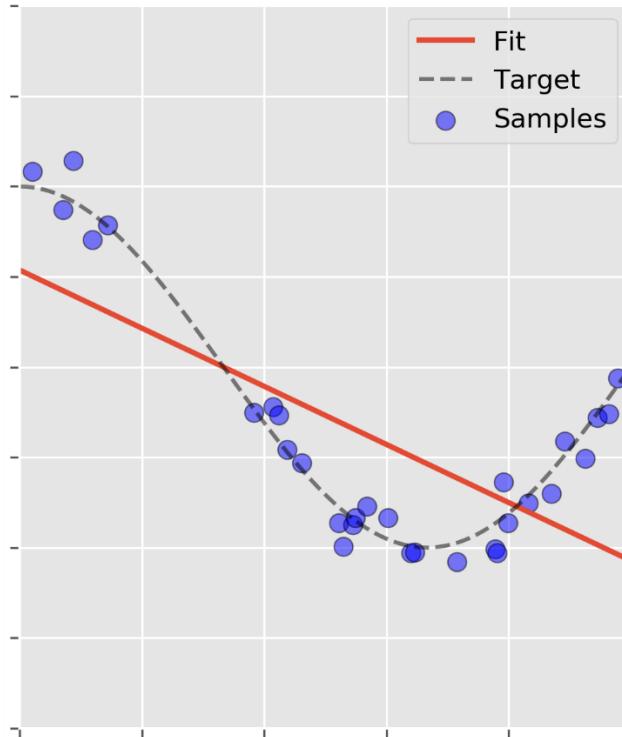


- بیشبرازش. یک مشکل بسیار متداول در یادگیری ماشین
 - مدل بیش از حد نیاز پیچیده
 - مثلا به دلیل تعداد بسیار زیاد ویژگی‌ها
 - عملکرد بسیار خوب مدل بر روی داده‌های آموزشی
 - عملکرد بسیار بد مدل بر روی داده‌های جدید
 - عدم قابلیت تعمیم برای داده‌های جدید!
- تنظیم. یک روش مؤثر برای کاهش یا حذف بیشبرازش.

بیشبرازش و اگریوشن

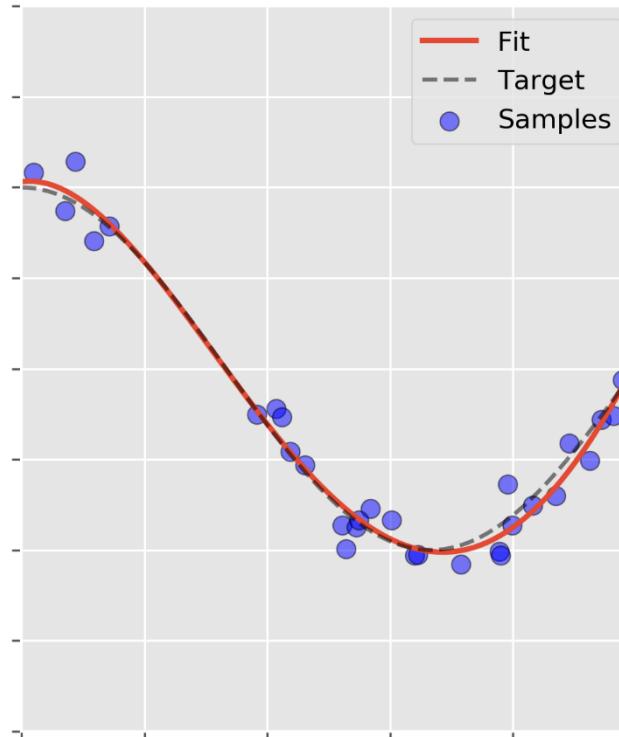
۲

Degree 1, MSE = 0.41



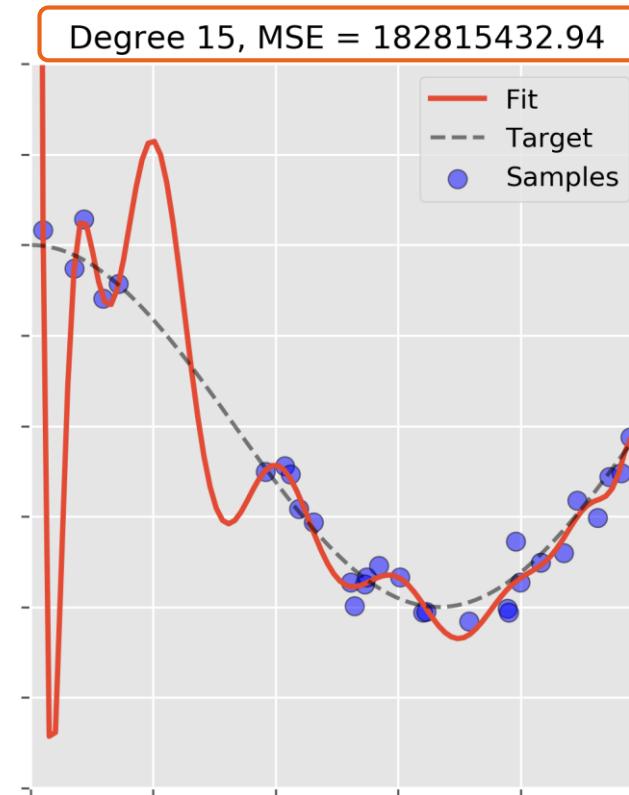
کم برآش

Degree 4, MSE = 0.04



مدل درست

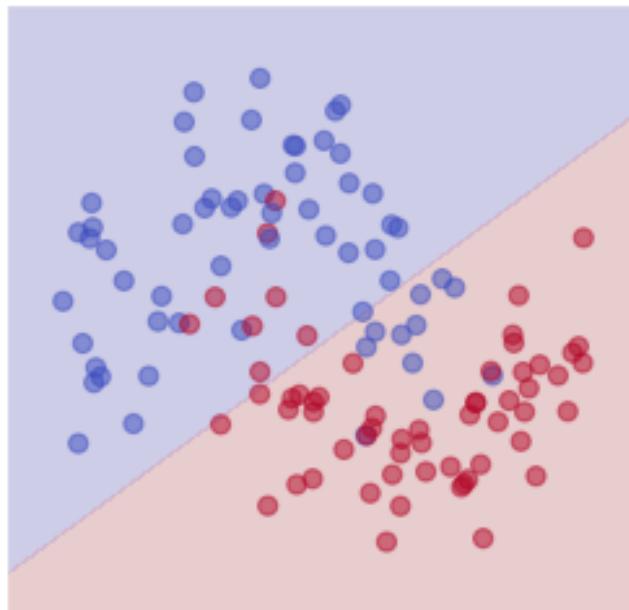
Degree 15, MSE = 182815432.94



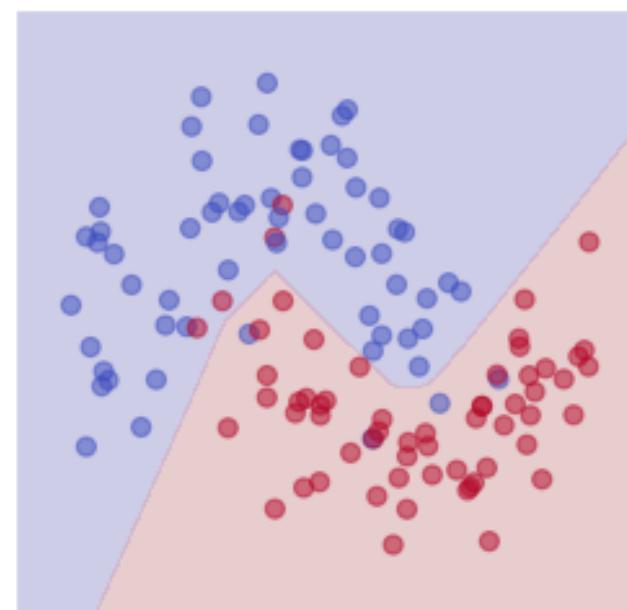
بیشبرآش

بیشبرازش و دسته‌بندی

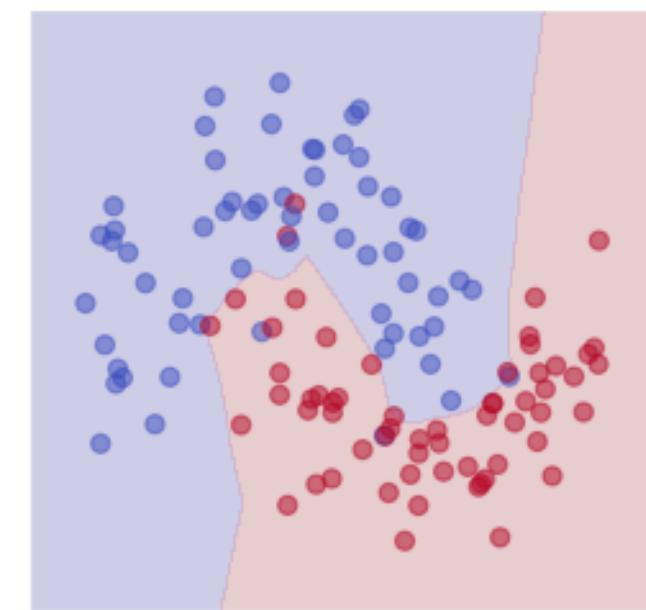
۴



کم‌برازش



مدل درست

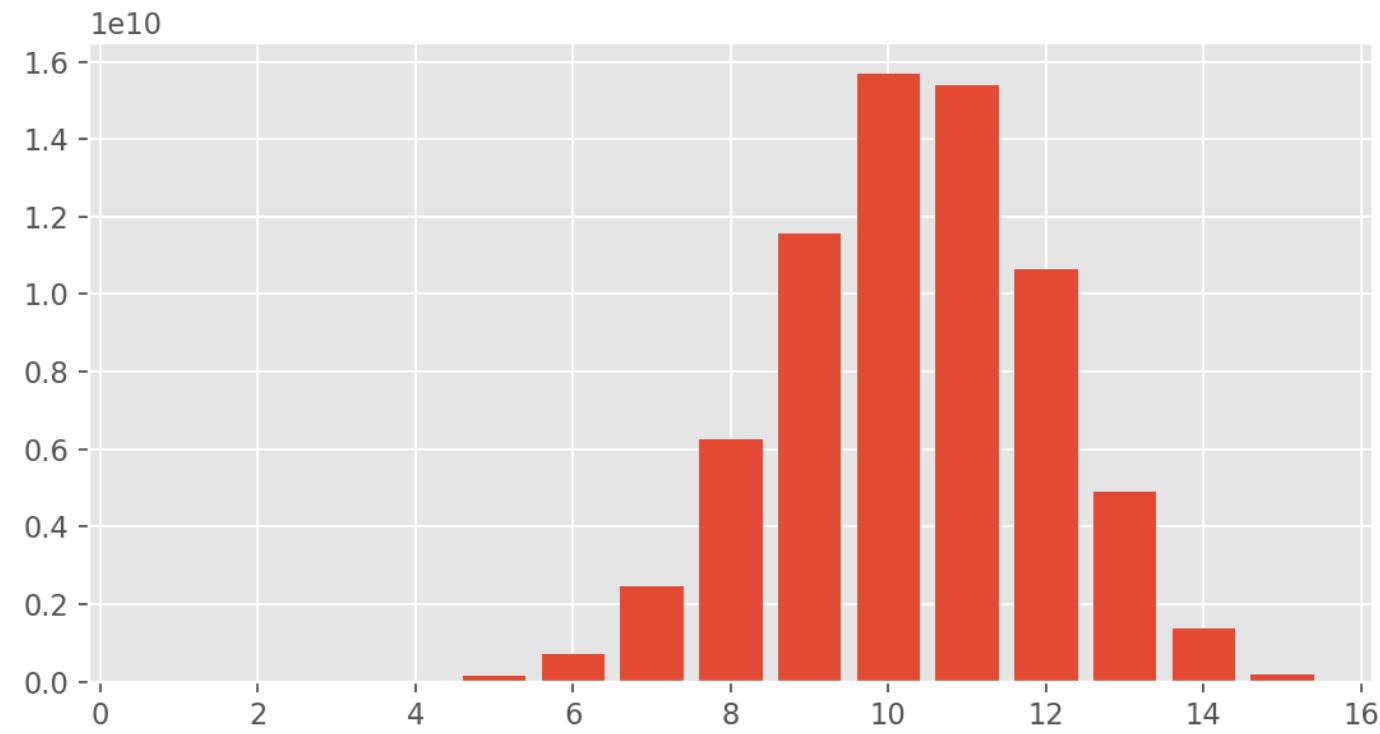
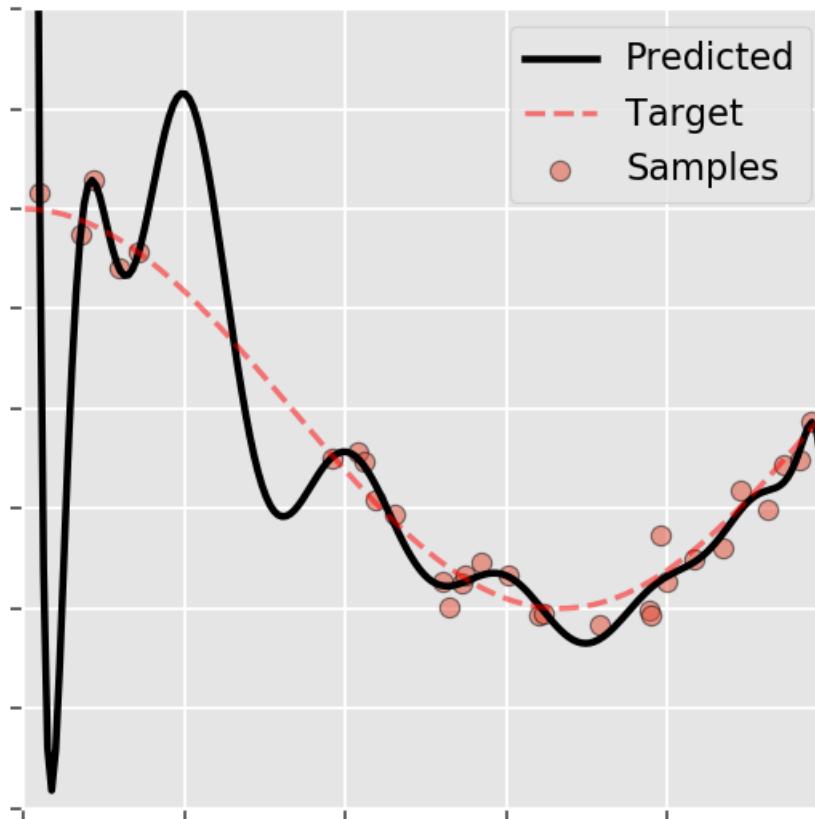


بیش‌برازش

تنظیم

۵

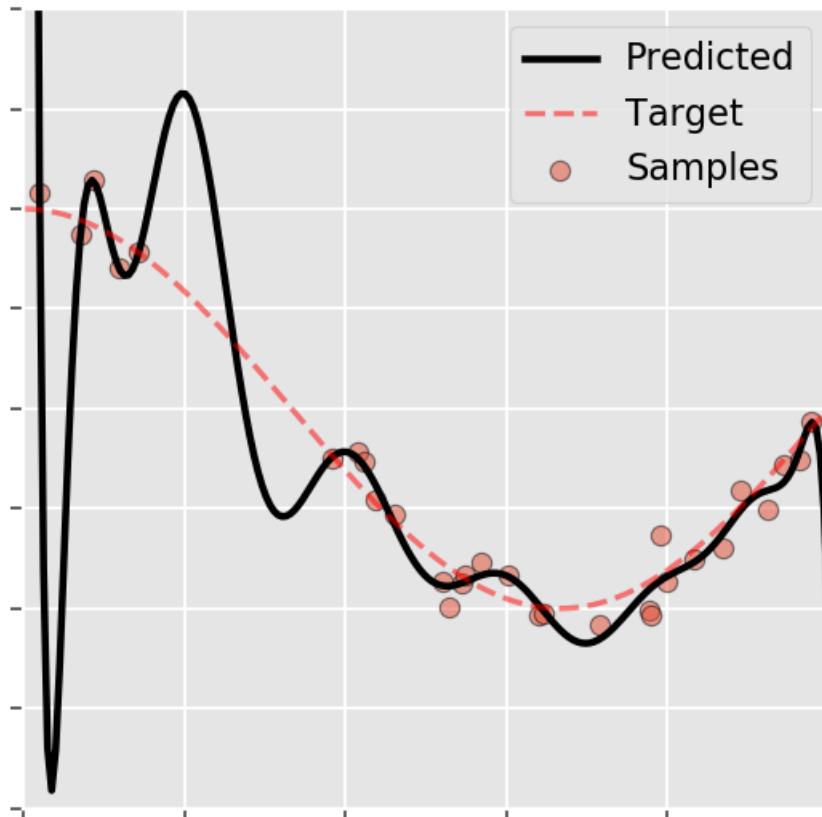
□ ایده. جلوگیری از بزرگ شدن بیش از حد مقدار پارامترها با افزودن یک جمله به تابع هزینه به منظور جریمه کردن مقادیر بزرگ پارامترها.



تنظیم

۶

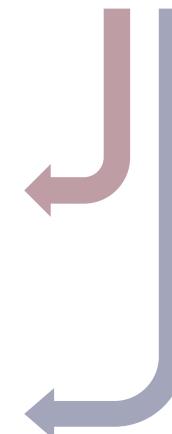
□ ایده. جلوگیری از بزرگ شدن بیش از حد مقدار پارامترها با افزودن یک جمله به تابع هزینه به منظور جریمه کردن مقادیر بزرگ پارامترها.



$$J(\theta) = \frac{1}{m} \sum_{i=1}^m Cost(x^{(i)}, y^{(i)}) + \lambda R(\theta)$$

$$R(\theta) = \sum_{j=1}^n \theta_j^2 = \|\theta\|_2^2 \quad \text{تنظیم L2}$$

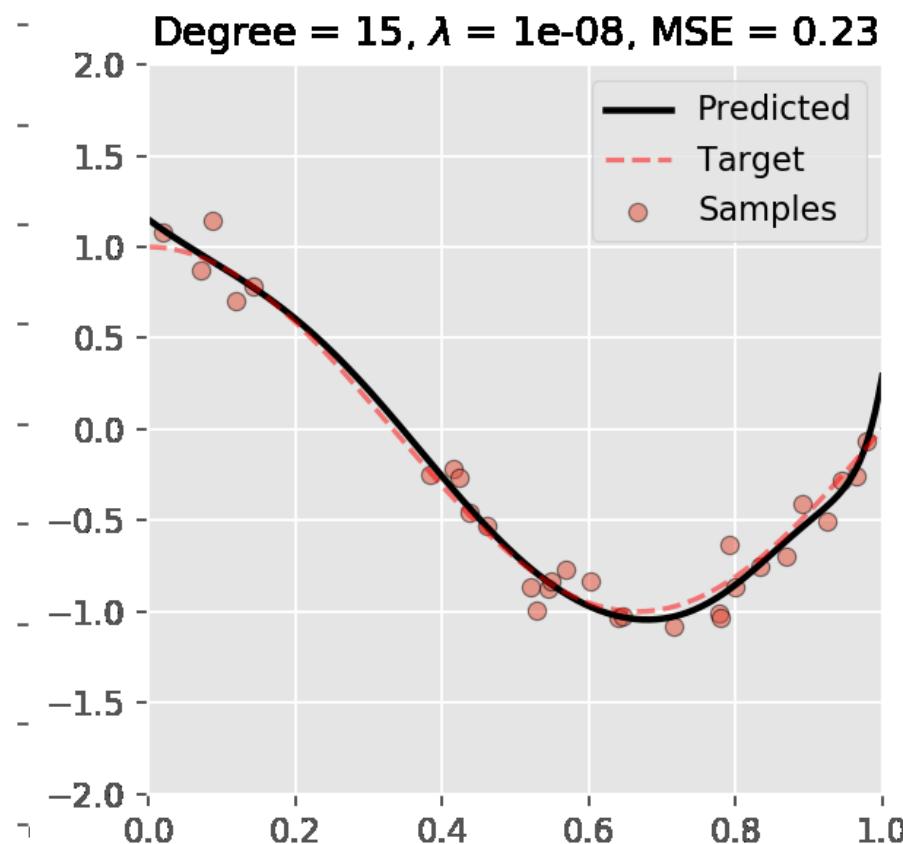
$$R(\theta) = \sum_{j=1}^n |\theta_j| = \|\theta\|_1 \quad \text{تنظیم L1}$$



تنظیم

۷

ایده. جلوگیری از بزرگ شدن بیش از حد مقدار پارامترها با افزودن یک جمله به تابع هزینه به منظور **جریمه کردن** مقادیر بزرگ پارامترها.



$$J(\theta) = \frac{1}{m} \sum_{i=1}^m Cost(x^{(i)}, y^{(i)}) + \lambda R(\theta)$$

ضریب تنظیم. برقراری توازن میان اهداف فوق.

دادن اهمیت بیشتر به فطاوی مجموعه آموزشی $\lambda \rightarrow 0$

دادن اهمیت بیشتر به فطاوی تعمیم $\lambda \rightarrow \infty$

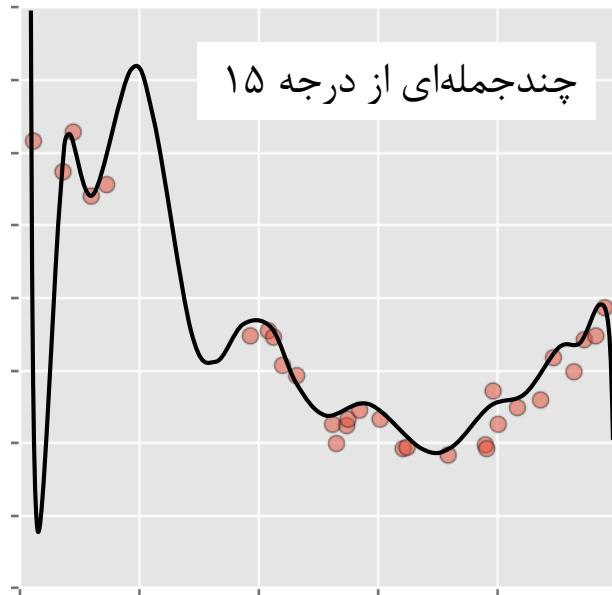
تابع هزينة



تنظیم

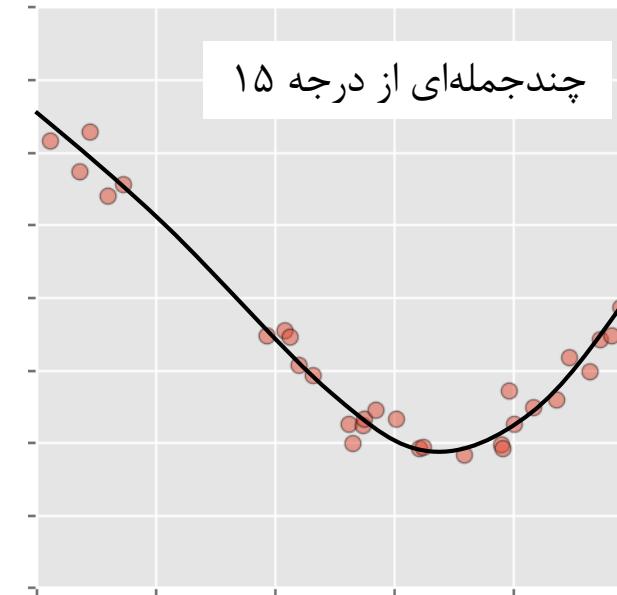
۹

رگرسیون بدون تنظیم



$$J(\theta) = \frac{1}{m} \sum_{i=1}^m Cost(x^{(i)}, y^{(i)})$$

رگرسیون با تنظیم

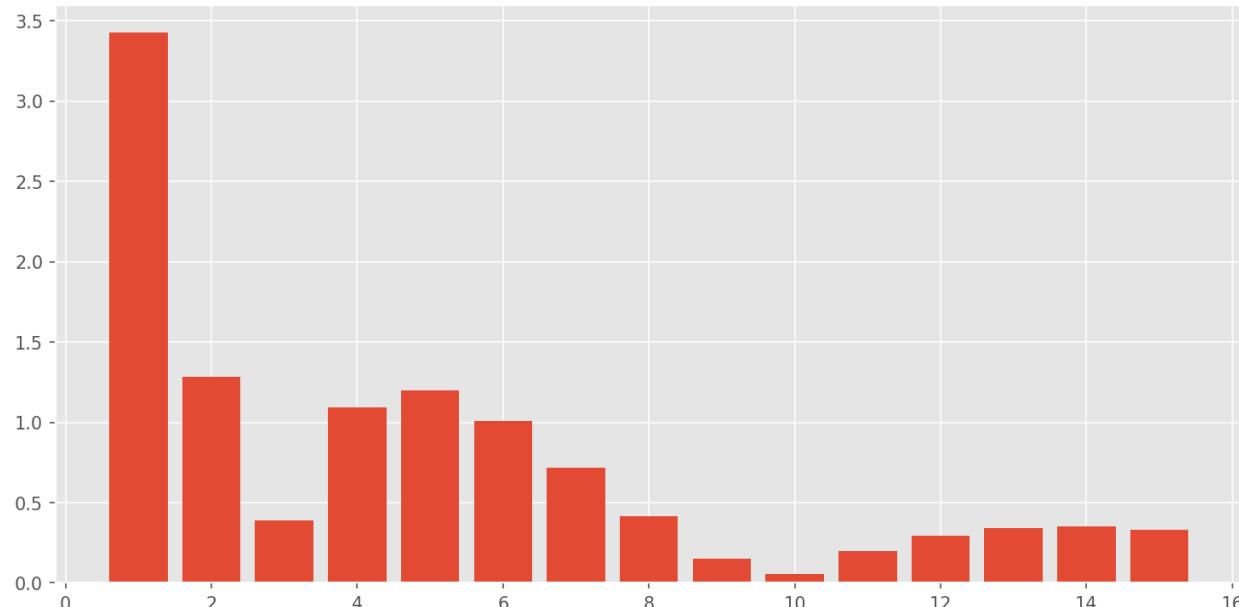


$$\frac{1}{m} \sum_{i=1}^m Cost(x^{(i)}, y^{(i)}) + \lambda R(\theta)$$

تنظیم

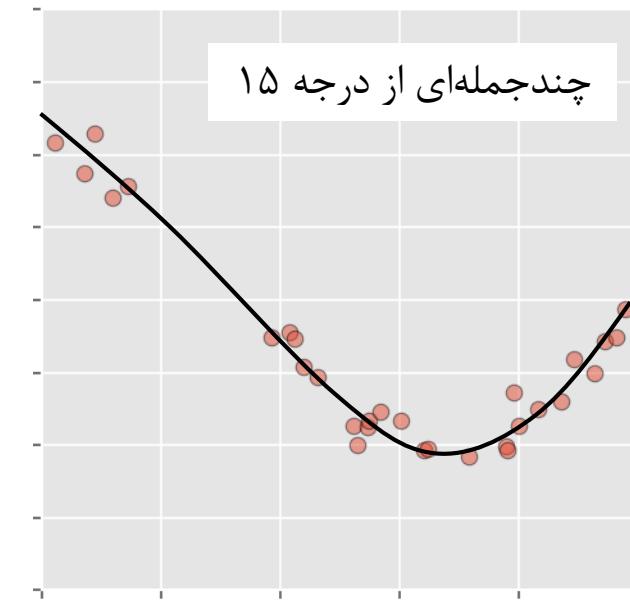
۱۰

مقدار پارامترها در صورت استفاده از تنظیم



$$R(\theta) = \sum_{j=1}^n \theta_j^2 = \|\theta\|_2^2 \quad \text{L2 تنظیم}$$

رگرسیون با تنظیم

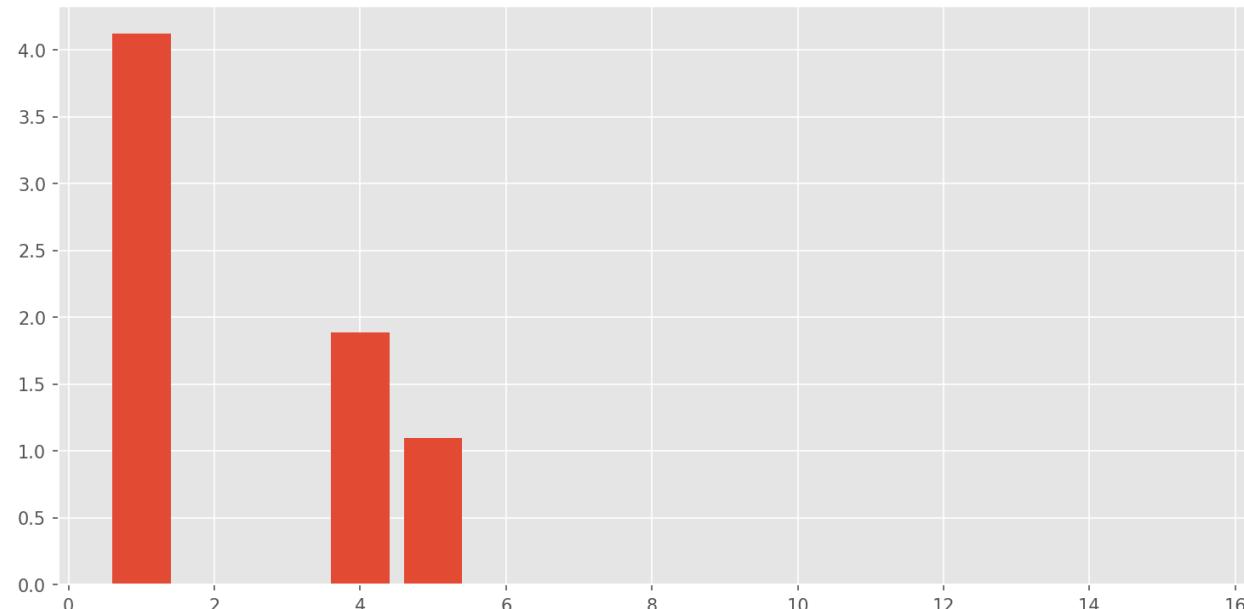


$$\frac{1}{m} \sum_{i=1}^m Cost(x^{(i)}, y^{(i)}) + \lambda R(\theta)$$

تنظیم

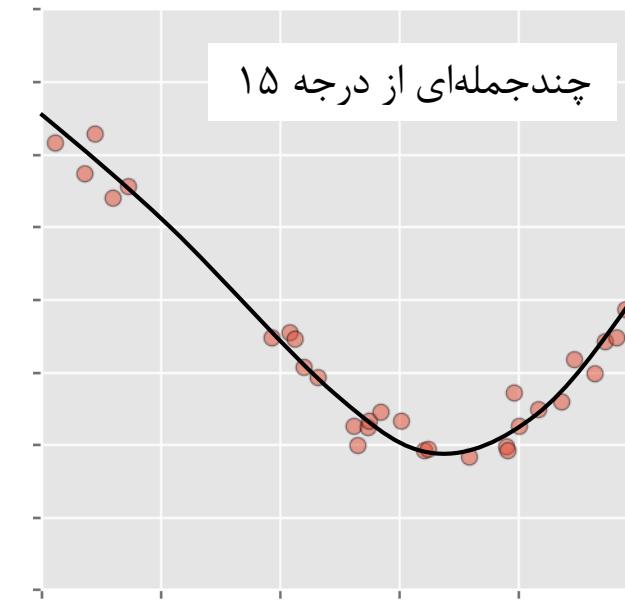
۱۱

مقدار پارامترها در صورت استفاده از تنظیم



$$R(\theta) = \sum_{j=1}^n |\theta_j| = \|\theta\|_1 \quad \text{L1 تنظیم}$$

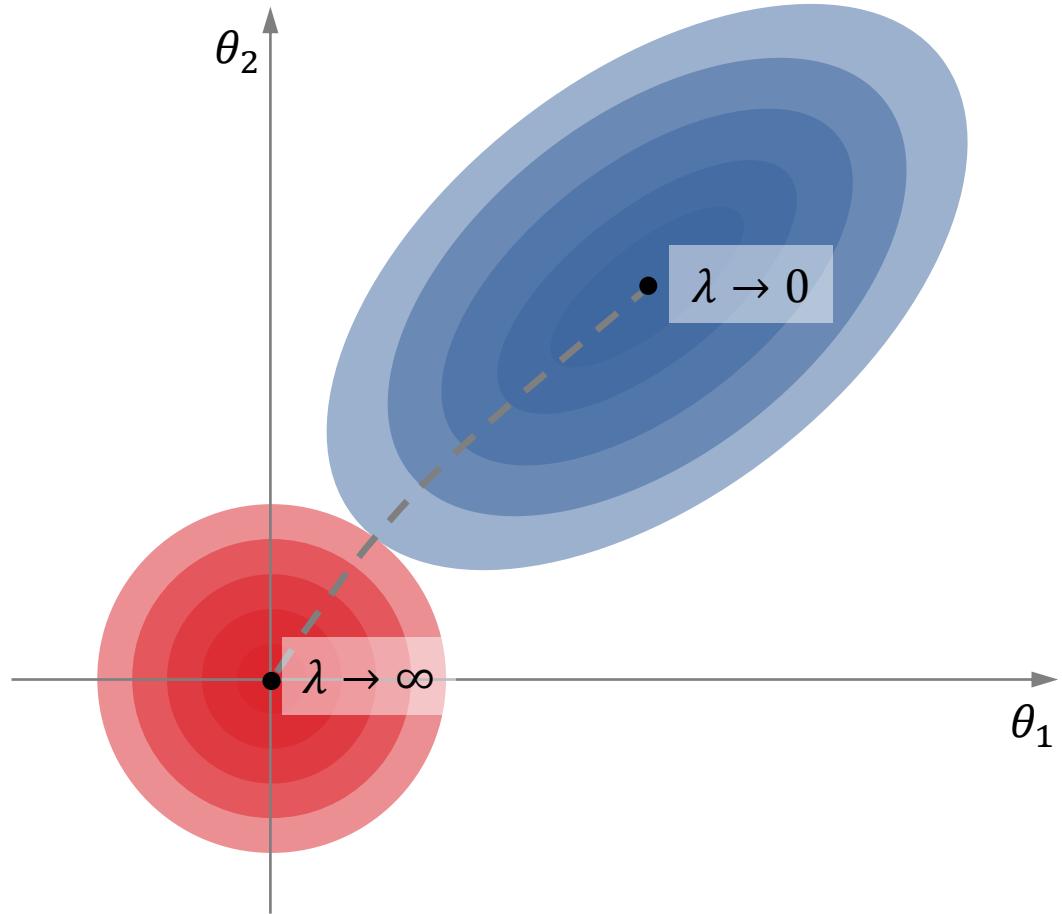
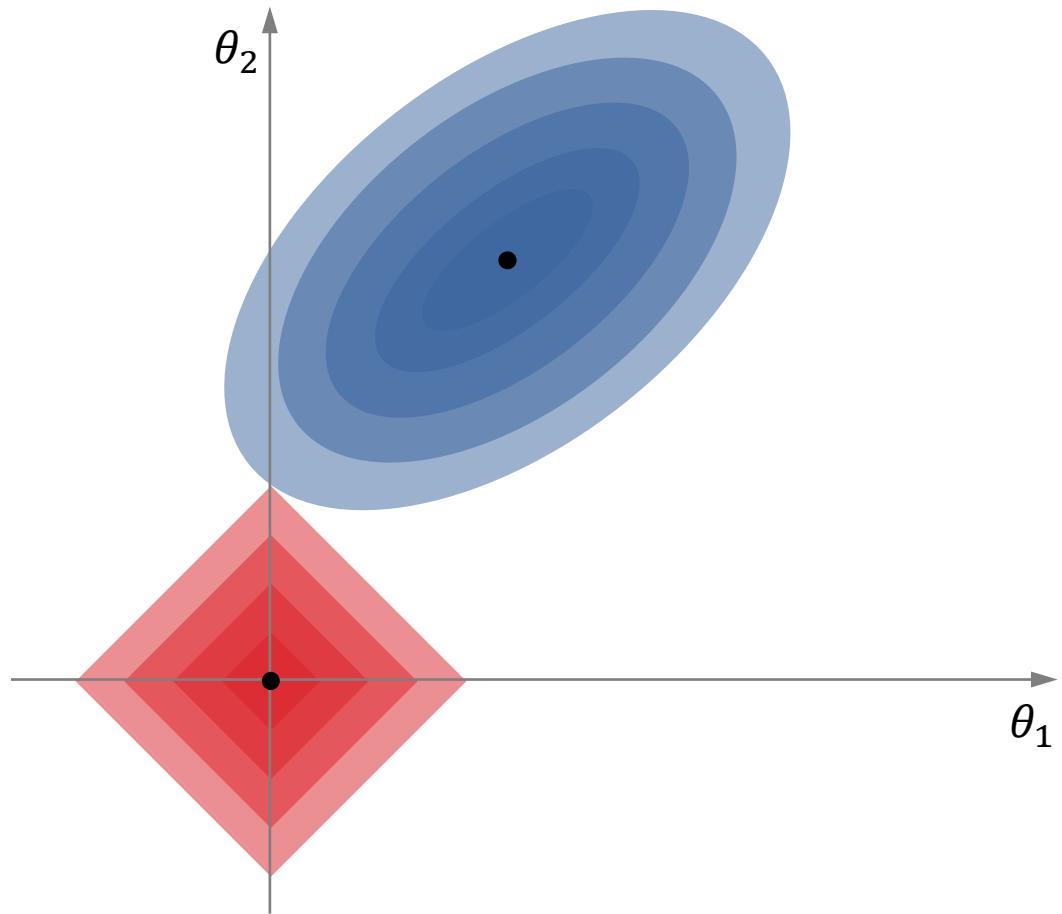
رگرسیون با تنظیم



$$\frac{1}{m} \sum_{i=1}^m Cost(x^{(i)}, y^{(i)}) + \lambda R(\theta)$$

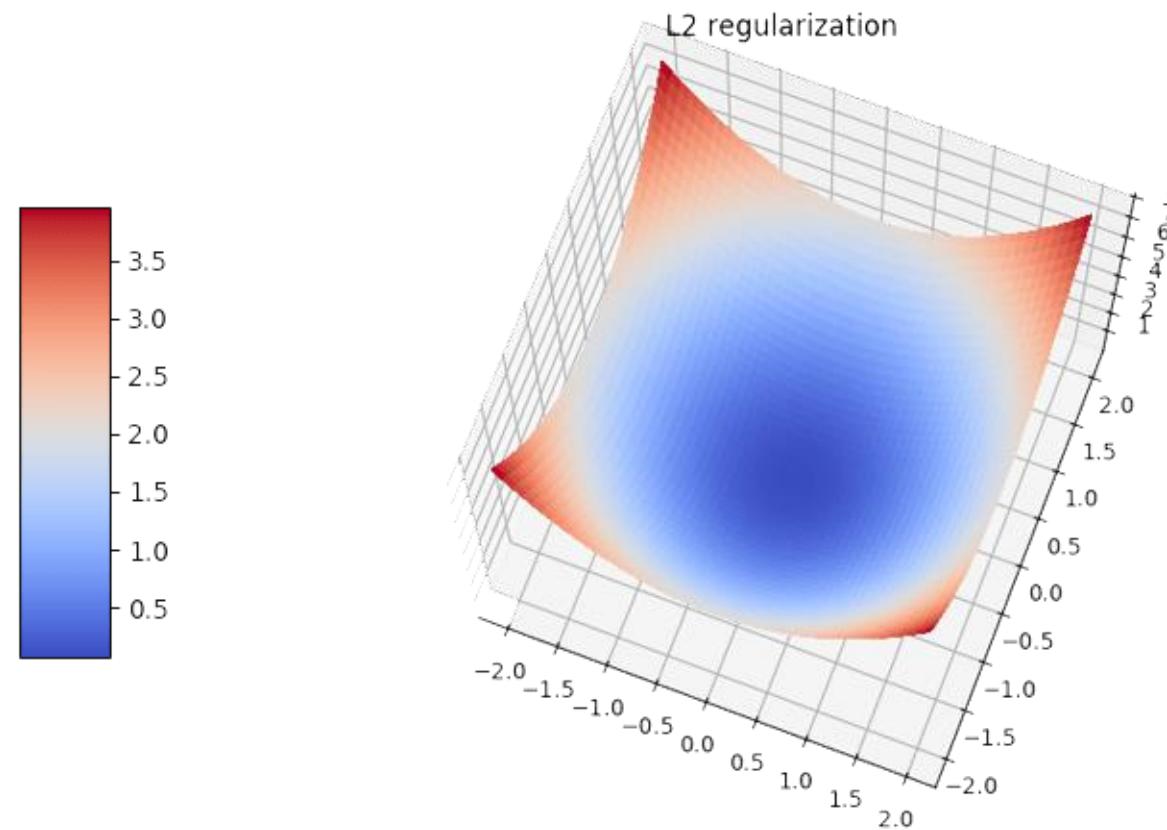
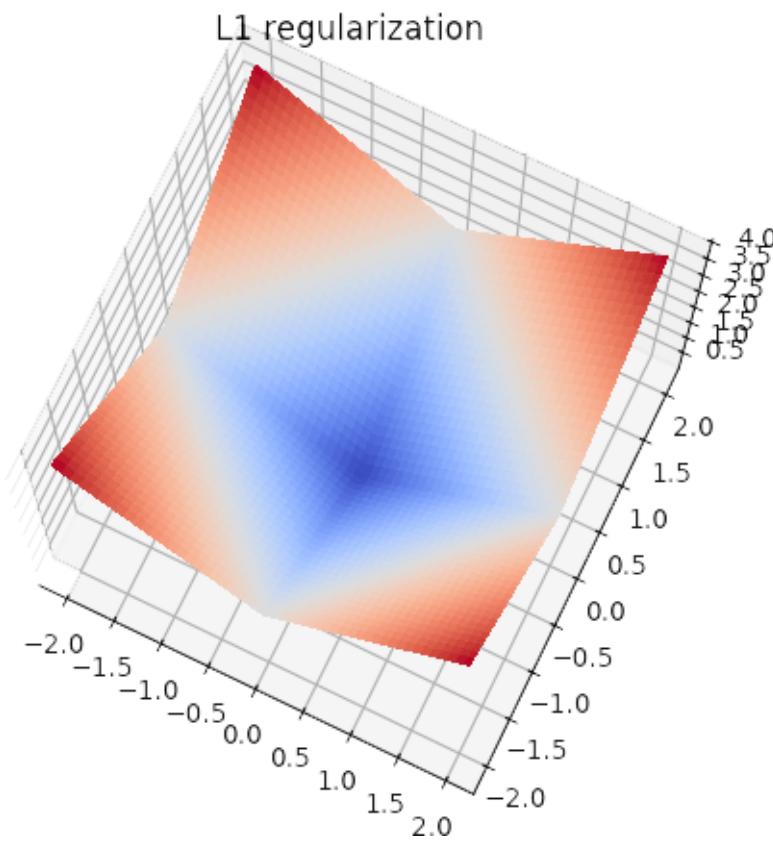
تنظیم L2 و L1

۱۲



تَنْظِيم L2 و L1

١٢



ڪرسيون خطي تنظيم شدہ

اگر سیون فطی تنظیم شده

۱۵

□ تابع هزینه.

$$J(\theta) = \frac{1}{2} \left[\sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right] = \frac{1}{2} (X\theta - y)^T (X\theta - y) + \frac{1}{2} \lambda \theta^T \theta$$

□ هدف. کمینه‌سازی تابع هزینه به منظور یافتن مقدار بهینه پارامترها

$$\min_{\theta} J(\theta)$$

گرادیان کاہشی (بدون تنظیم)

۱۶

□ بدون استفاده از تنظیم.

repeat until convergence {

$$\theta_j = \theta_j - \alpha \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)} \quad (j = 0, 1, 2, \dots, n)$$

}

گرادیان کاہشی (با تنظیم)

۱۷

با استفاده از تنظیم. □

repeat until convergence {

$$\theta_0 = \theta_0 - \alpha \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_0^{(i)}$$

$$\theta_j = \theta_j - \alpha \left[\sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)} + \lambda \theta_j \right] \quad (j = 1, 2, \dots, n)$$

}



$$\theta_j = \underbrace{\theta_j (1 - \alpha \lambda)}_{< 1} - \alpha \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}$$

مُعَادِل نِرْمَال (با تَنْظِيم)

۱۸

$$J(\theta) = \frac{1}{2} (X\theta - y)^T (X\theta - y) + \frac{1}{2} \lambda \theta^T \theta$$

$$\theta = \underbrace{(X^T X + \lambda I)^{-1}}_{(\lambda > 0)} X^T y$$

وارون پذیر

$$\frac{\partial J}{\partial \theta} = X^T (X\theta - y) + \lambda \theta$$

$$= X^T X \theta - X^T y + \lambda \theta$$

$$= (X^T X + \lambda I) \theta - X^T y = 0$$

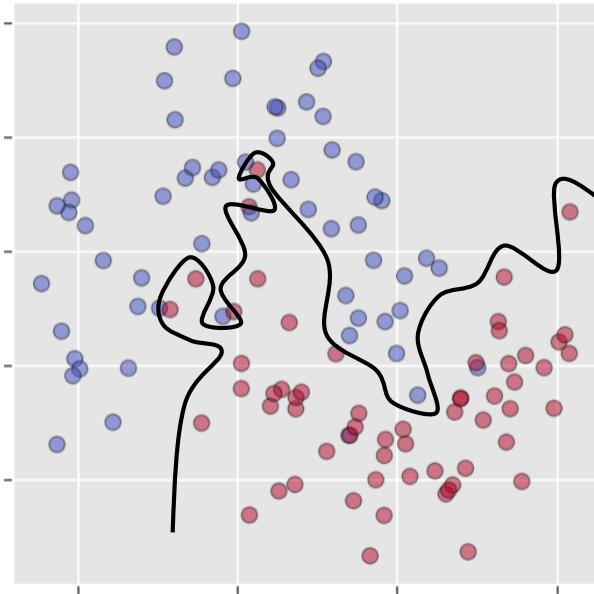
$$(X^T X + \lambda I) \theta = X^T y$$

$$\theta = \left(X^T X + \lambda \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \right)^{-1} X^T y$$

رگرسیون لجستیک تنظیم شده

گرسیون لجستیک (بدون تنظیم)

۲۰



چندجمله‌ای از درجه ۱۵

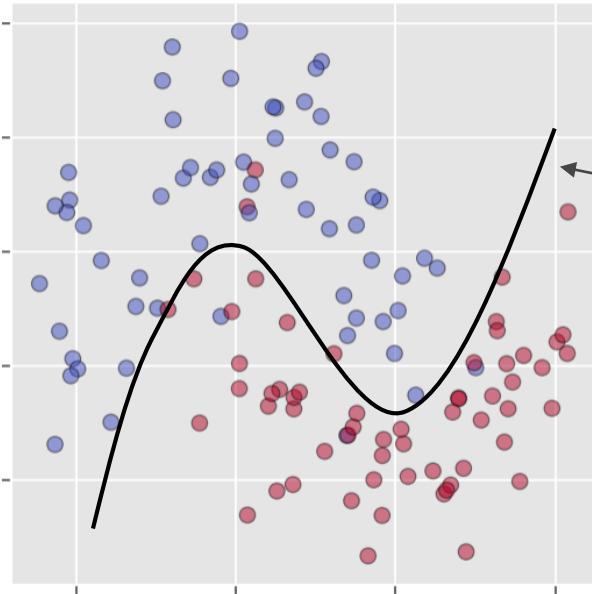
فرضیه. □

تابع هزینه. □

$$J(\theta) = - \sum_{i=1}^m y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))$$

گرسيون لجسيك (با تنظيم)

۲۱



چندجمله‌ای از درجه ۱۵

فرضيه. □

تابع هزينه. □

$$J(\theta) = - \sum_{i=1}^m y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))$$

$$+ \frac{\lambda}{2} \sum_{j=1}^n \theta_j^2$$

گرادیان کاہشی

۲۲

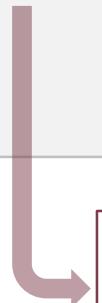
با استفاده از تنظیم. □

repeat until convergence {

$$\theta_0 = \theta_0 - \alpha \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_0^{(i)}$$

$$\theta_j = \theta_j - \alpha \left[\sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)} + \lambda \theta_j \right] \quad (j = 1, 2, \dots, n)$$

}



$$h_\theta(x^{(i)}) = \frac{1}{1 + e^{-\theta^T x^{(i)}}}$$

برهینه‌سازی پیش‌رفته

۲۲

```
from scipy.optimize import minimize  
  
minimize(J, x0, method='CG', jac=grads)
```

$$\sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)} + \lambda \theta_j$$
$$-\sum_{i=1}^m y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)})) + \frac{\lambda}{2} \sum_{j=1}^n \theta_j^2$$

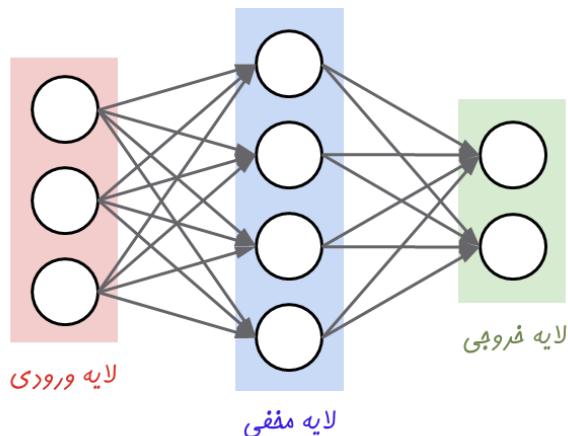
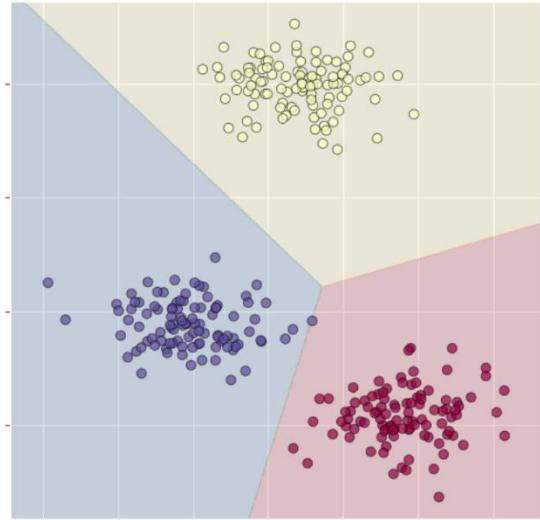
شبکه‌های عصبی مصنوعی

سید ناصر رضوی www.snrazavi.ir

۱۳۹۷

فهرست

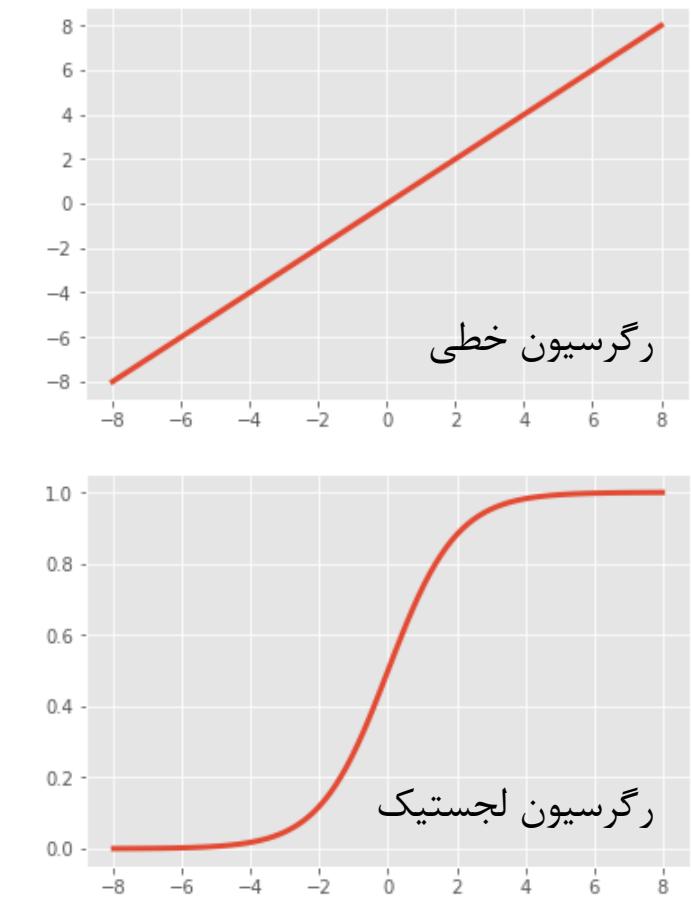
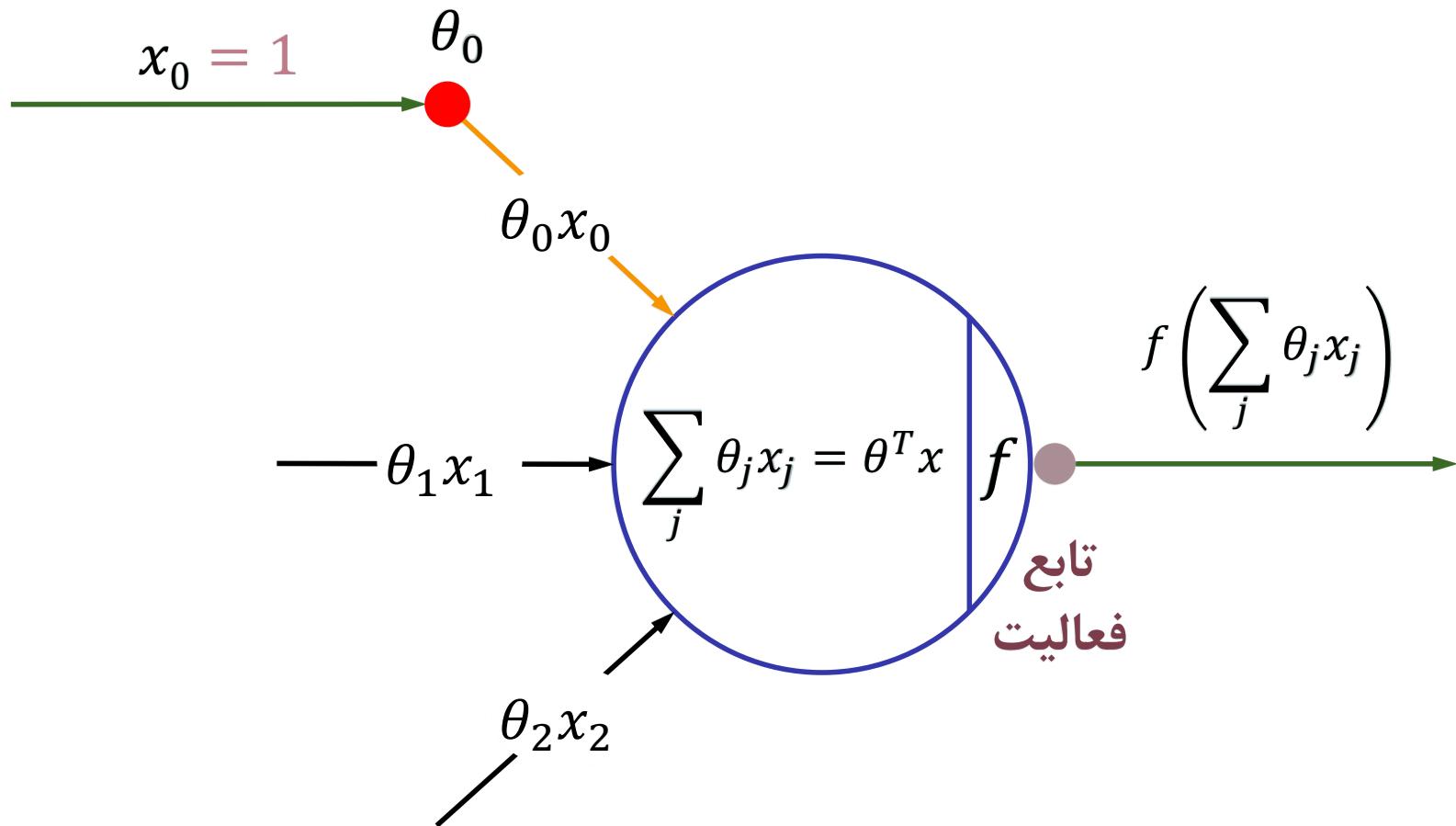
۲



- یادآوری رگرسیون لجستیک.
- رگرسیون لجستیک چند دسته‌ای (چند کلاسی)
 - دسته‌بندی یکی در برابر بقیه
- دسته‌بند سافت‌مکس.
- تابع هزینه سافت‌مکس
- آموزش دسته‌بند سافت‌مکس و گرادیان کاهشی
 - تفسیر هندسی
- شبکه‌های عصبی.
- مرحله انتشار پیش‌رو
- مرحله انتشار پس‌رو

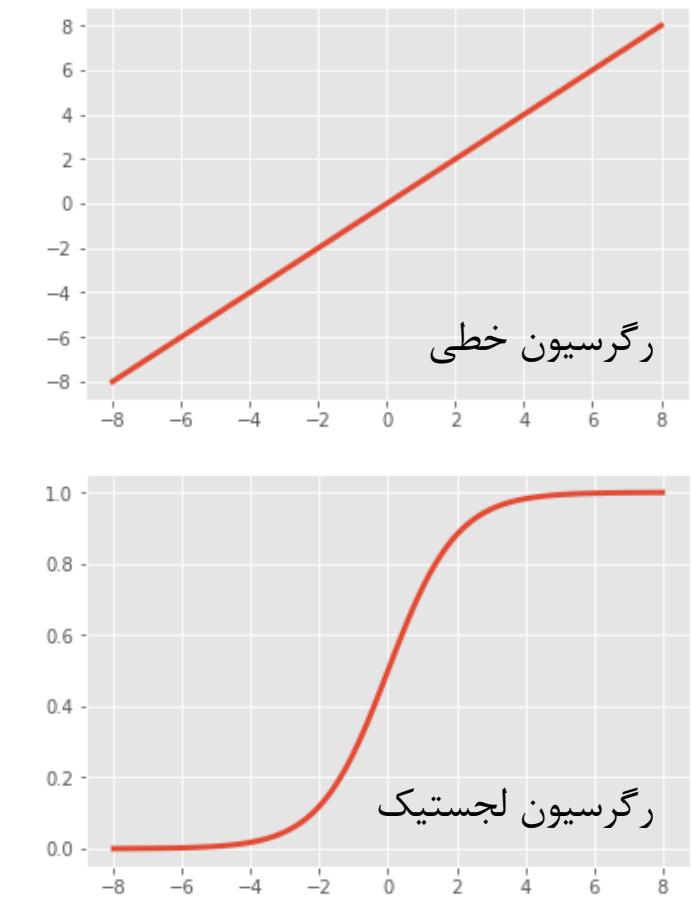
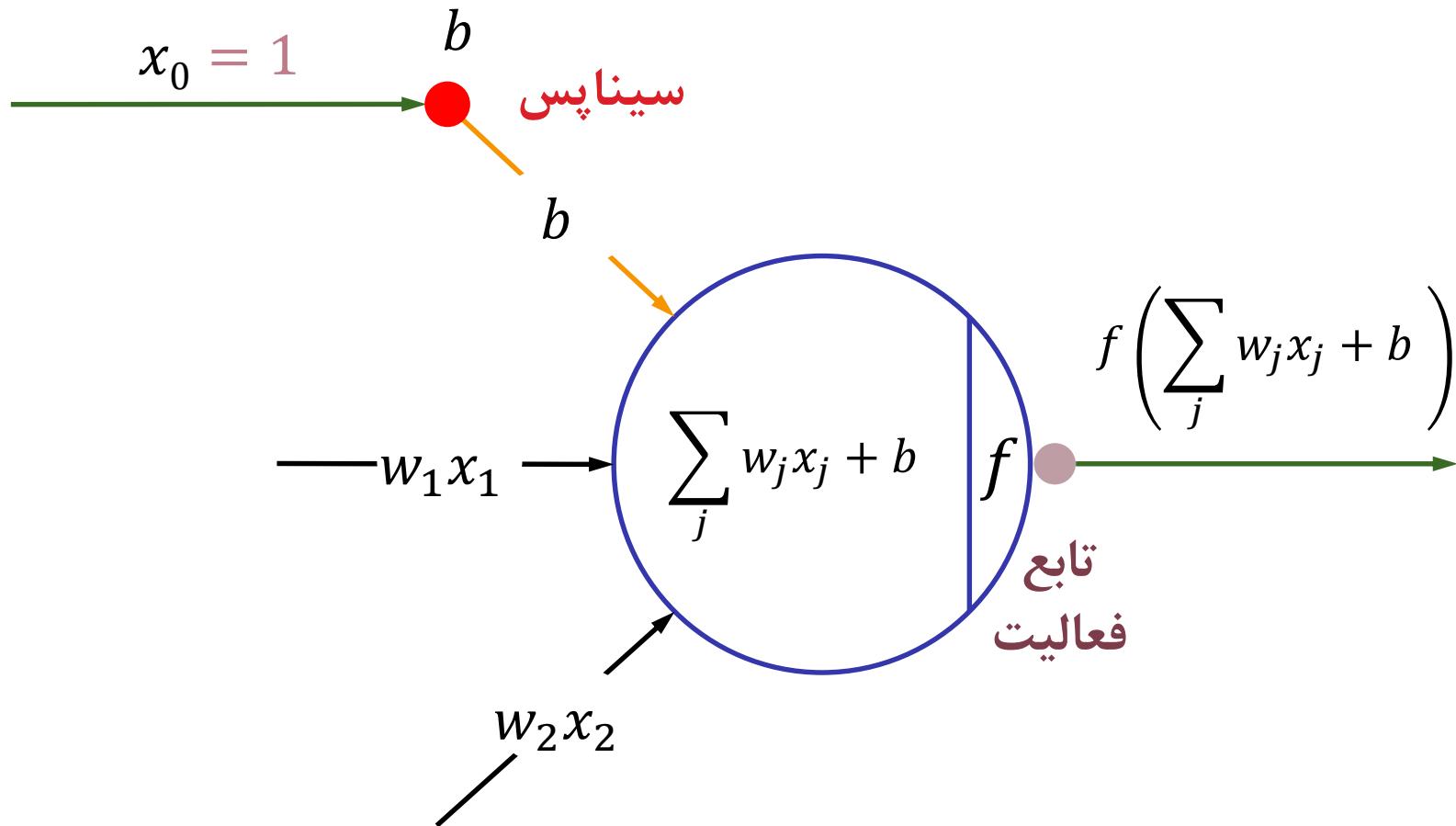
یادآوری: رگرسیون لجستیک دودویی

۳



یادآوری: رگرسیون لجستیک دودویی

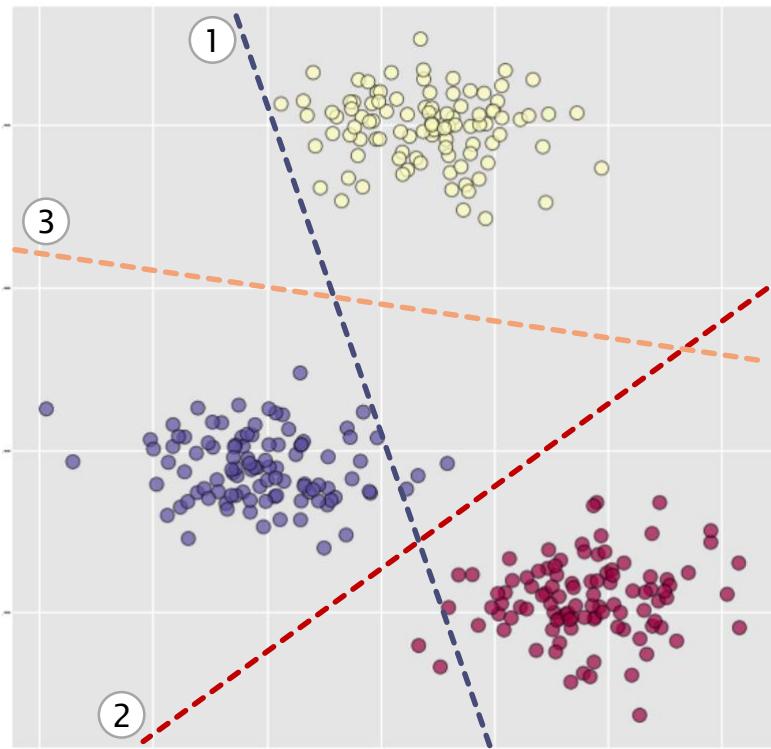
۴



یادآوری: رگرسیون لجستیک پندر دسته‌ای

۵

یکی در برابر بقیه. ایجاد یک دسته‌بند به ازای هر دسته. □



$$h^{(1)}(x) = g\left(\left(\theta^{(1)}\right)^T x\right)$$

$$h^{(2)}(x) = g\left(\left(\theta^{(2)}\right)^T x\right)$$

$$h^{(3)}(x) = g\left(\left(\theta^{(3)}\right)^T x\right)$$

دسته‌بندی داده جدید. □

$$y = \arg \max_c h^{(c)}(x)$$

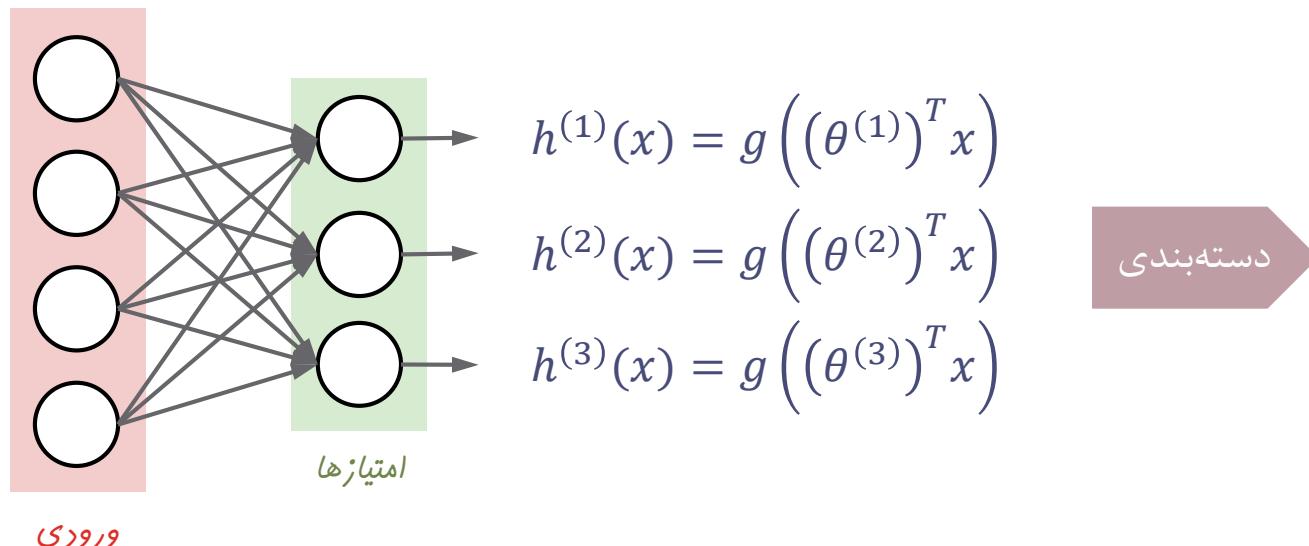
یادآوری: رگرسیون لجستیک پندهای دسته‌ای

۶

□ یکی در برابر بقیه. ایجاد یک دسته‌بند به ازای هر دسته.

□ دسته‌بندی داده جدید.

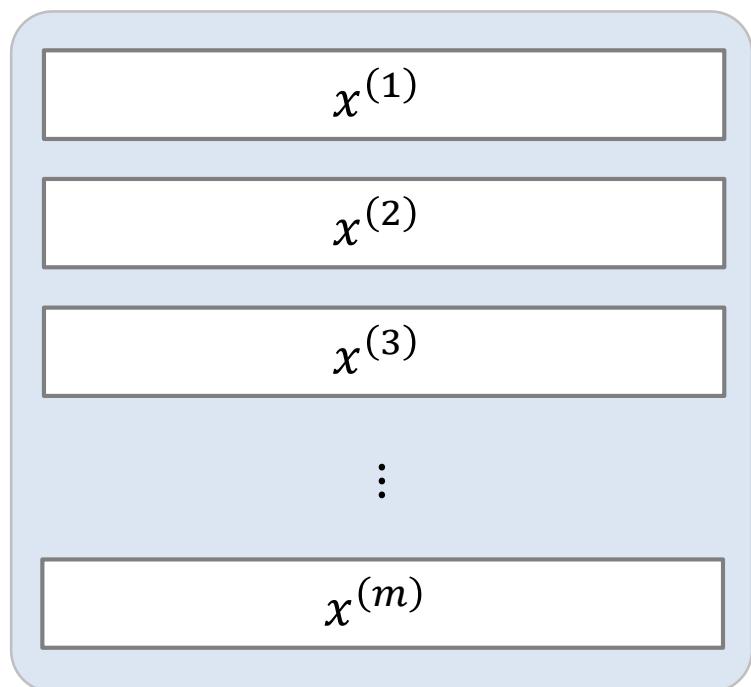
□ مثال. چهار ویژگی و سه دسته.



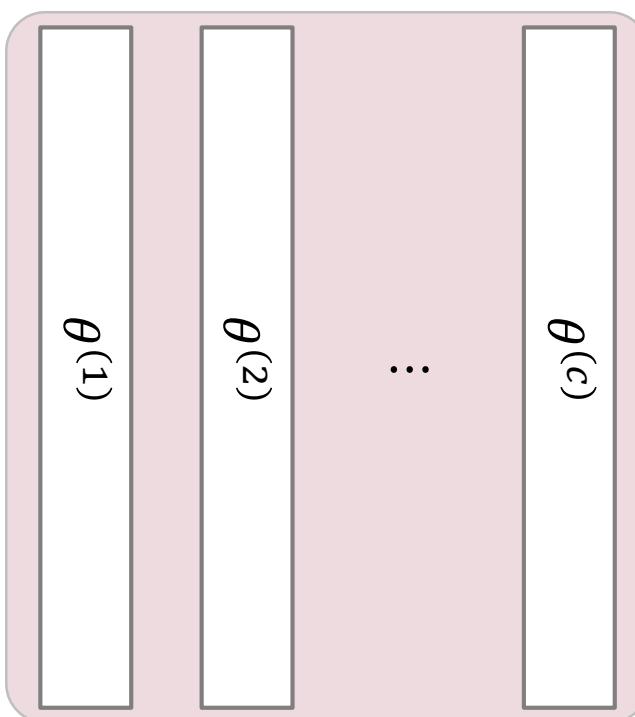
کریشن لجستیک پنداشتهای: بوداگریسازی

v

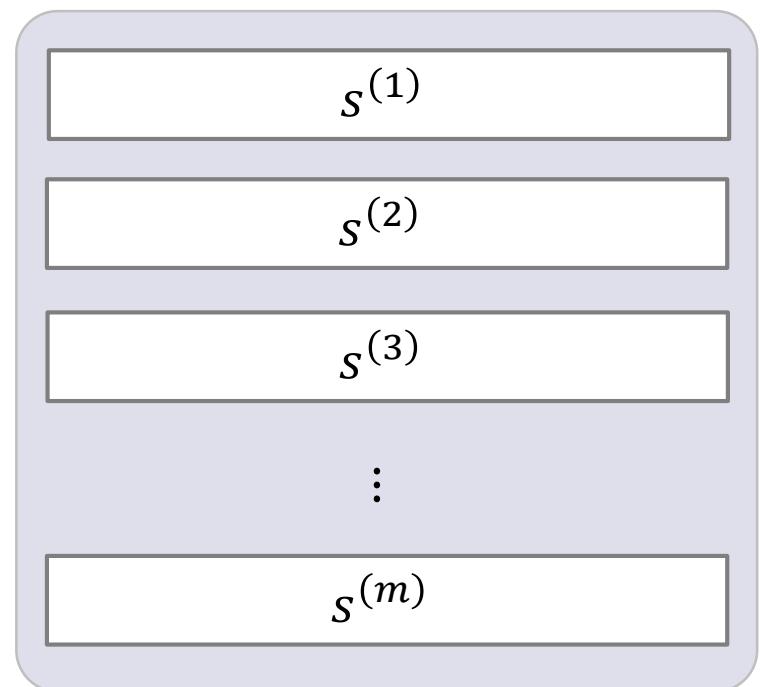
$X_{m \times n}$



$\theta_{n \times c}$



scores



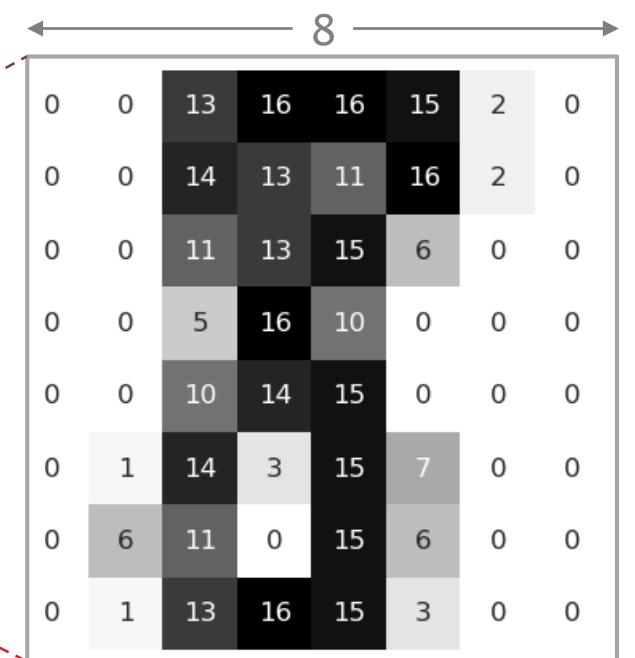
scores = $X @ \Theta + \theta_0$

رگرسیون لجستیک پنده دسته‌ای: تشفیض ارقام دسته‌نویس

۸

۱۰ دسته و ۶۴ ویژگی

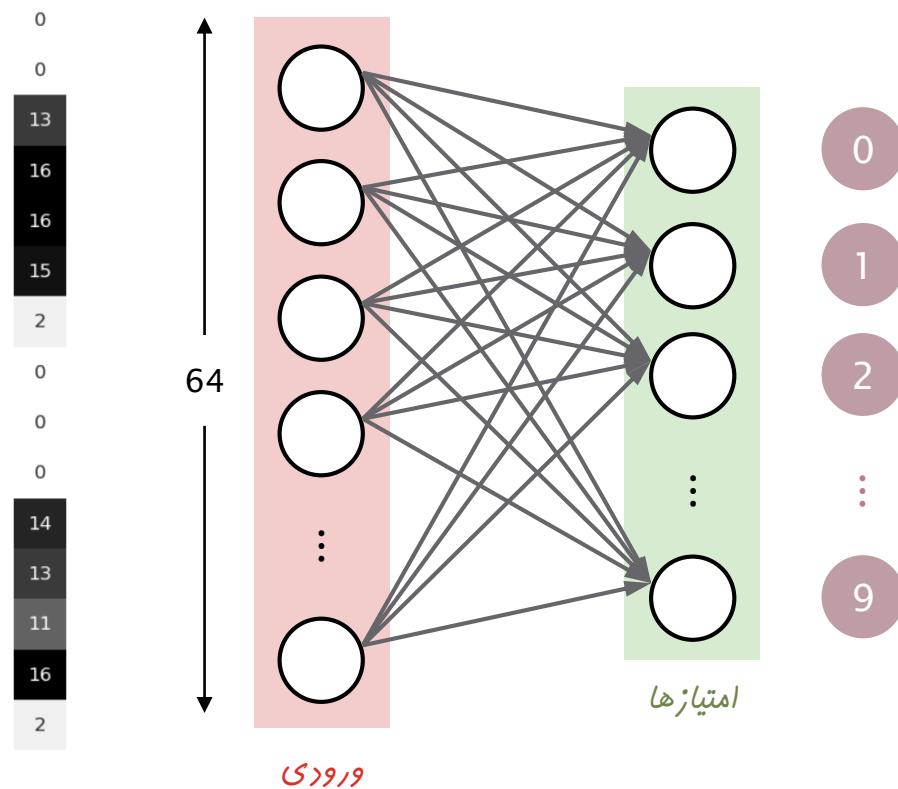
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹



رگرسیون لجستیک پنده دسته‌ای: تشفیض ارقام دسته‌نویس

۹

۱۰ دسته و ۶۴ ویژگی



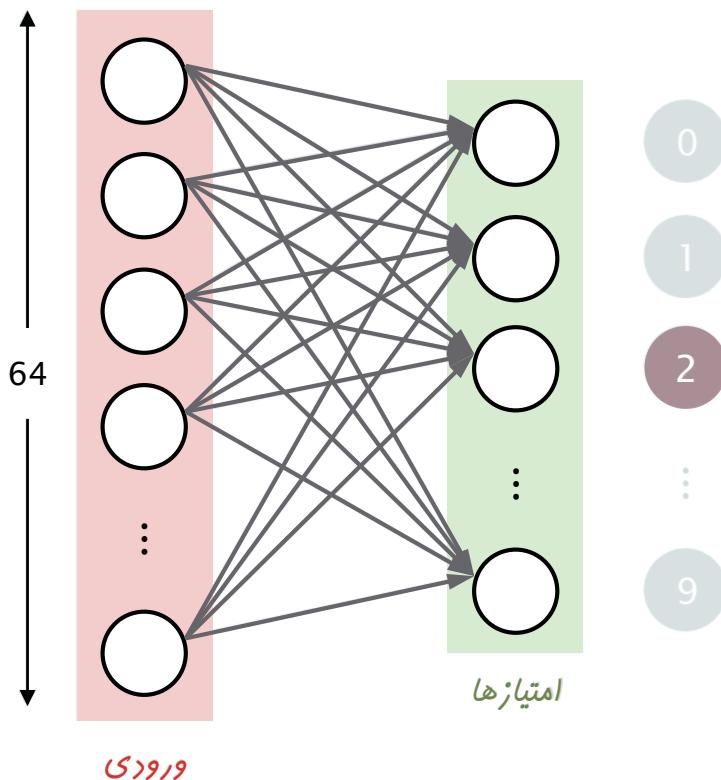
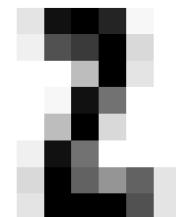
An 8x8 grid of numbers representing a digit image. The grid is labeled with dimensions 8x8. The values in the grid are as follows:

0	0	13	16	16	15	2	0
0	0	14	13	11	16	2	0
0	0	11	13	15	6	0	0
0	0	5	16	10	0	0	0
0	0	10	14	15	0	0	0
0	1	14	3	15	7	0	0
0	6	11	0	15	6	0	0
0	1	13	16	15	3	0	0

رگرسیون لجستیک چند دسته‌ای: پیش‌بینی

۱۰

□ تفسیر هندسی. محاسبه شباهت بردار ورودی با بردارهای وزن مربوط به تک‌تک دسته‌ها.



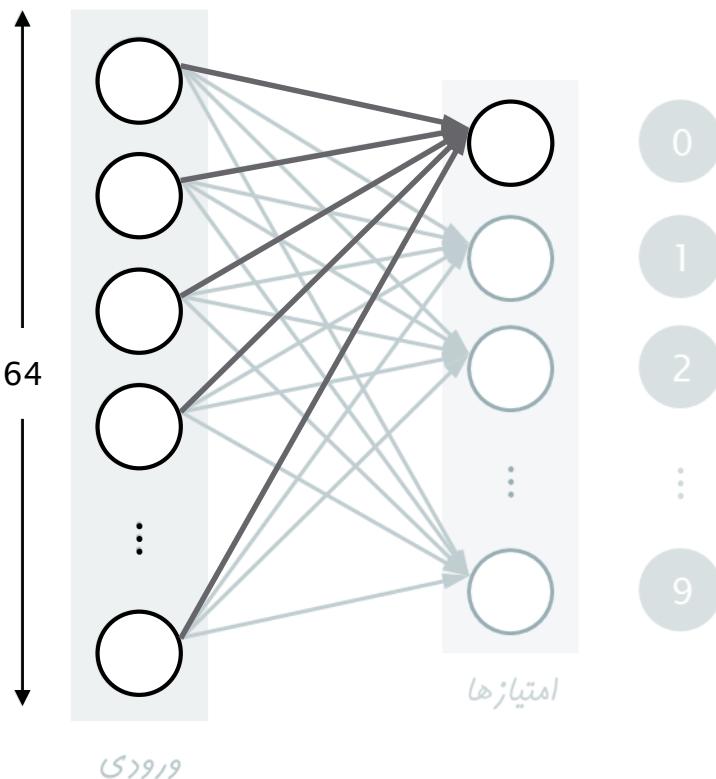
0
1
2
⋮
9

```
def predict(W, b, X):  
    scores = X @ W + b  
    return np.argmax(scores, axis=1)
```

رگرسیون لجستیک چند دسته‌ای: پیش‌بینی

۱۱

محاسبه شباهت بردار ورودی با
پارامترهای مربوط به کلاس **صفحه**

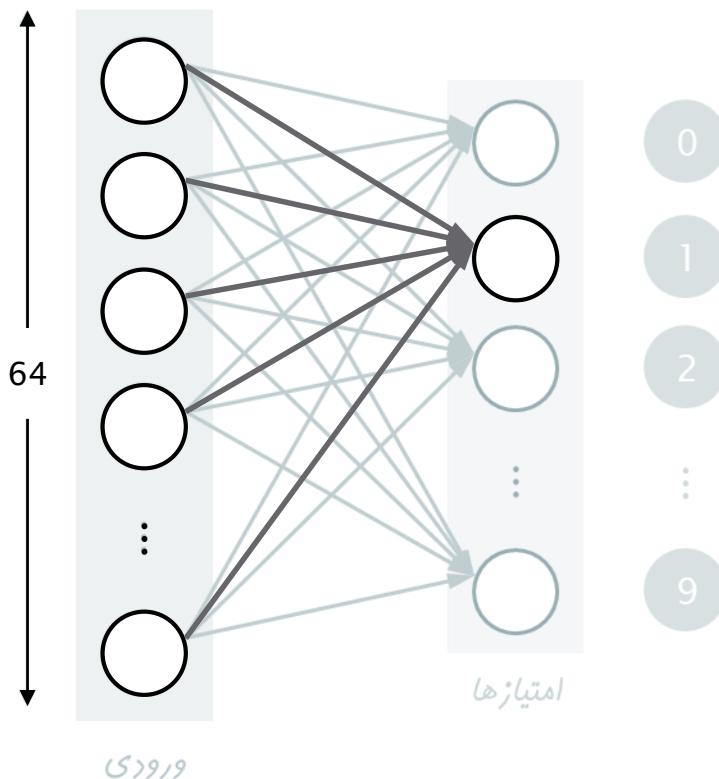
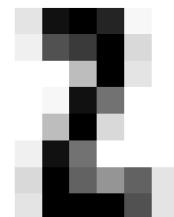


```
def predict(W, b, X):  
    scores = X @ W + b  
    return np.argmax(scores, axis=1)
```

رگرسیون لجستیک چند دسته‌ای: پیش‌بینی

۱۲

محاسبه شباهت بردار ورودی با
پارامترهای مربوط به کلاس یک

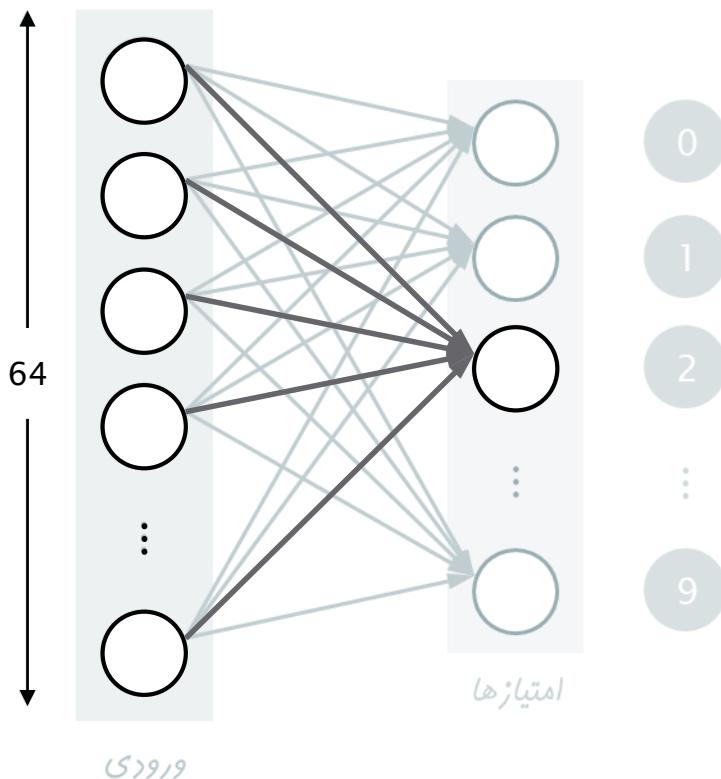
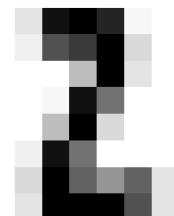


```
def predict(W, b, X):  
    scores = X @ W + b  
    return np.argmax(scores, axis=1)
```

رگرسیون لجستیک چند دسته‌ای: پیش‌بینی

۱۲

محاسبه شباهت بردار ورودی با
پارامترهای مربوط به کلاس \hat{w}

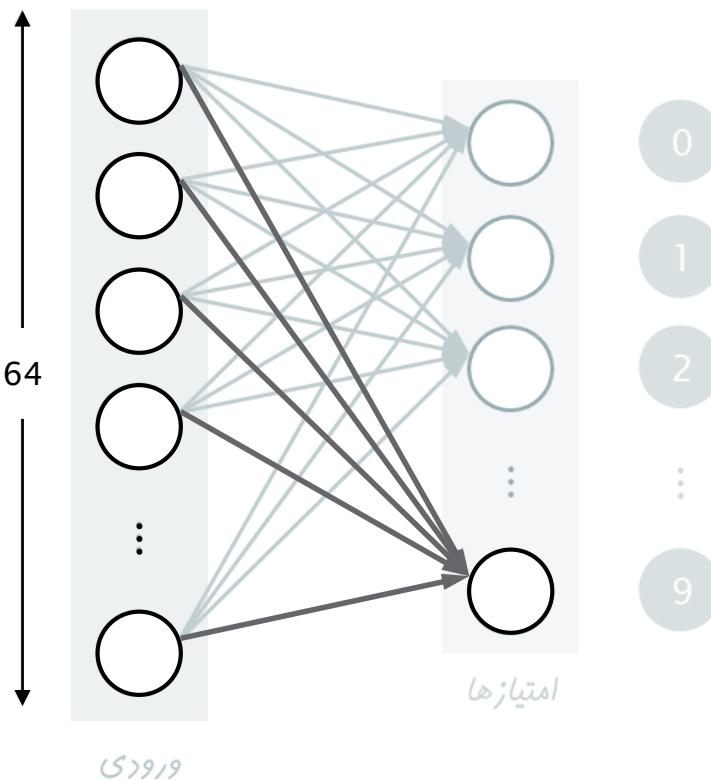


```
def predict(W, b, X):  
    scores = X @ W + b  
    return np.argmax(scores, axis=1)
```

رگرسیون لجستیک چند دسته‌ای: پیش‌بینی

۱۴

محاسبه شباهت بردار ورودی با
پارامترهای مربوط به کلاس \hat{y}



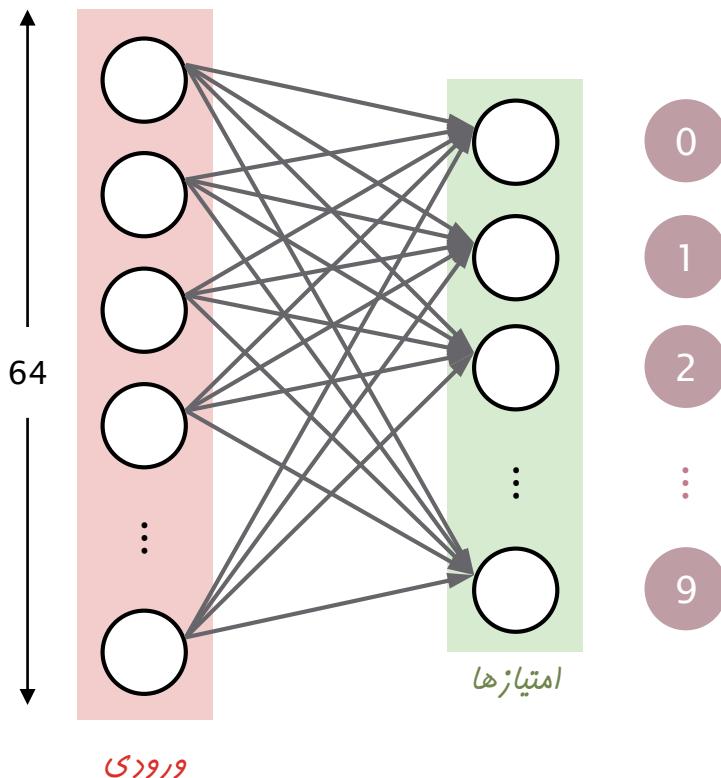
```
def predict(W, b, X):  
    scores = X @ W + b  
    return np.argmax(scores, axis=1)
```

دمسَبند سافت مکس

اگر سیون لجستیک دودویی: تابع سیگموید

۱۶

□ دسته‌بندی دودویی. مقدار تابع فرضیه بیانگر احتمال تعلق داده ورودی به دسته ۱ است.



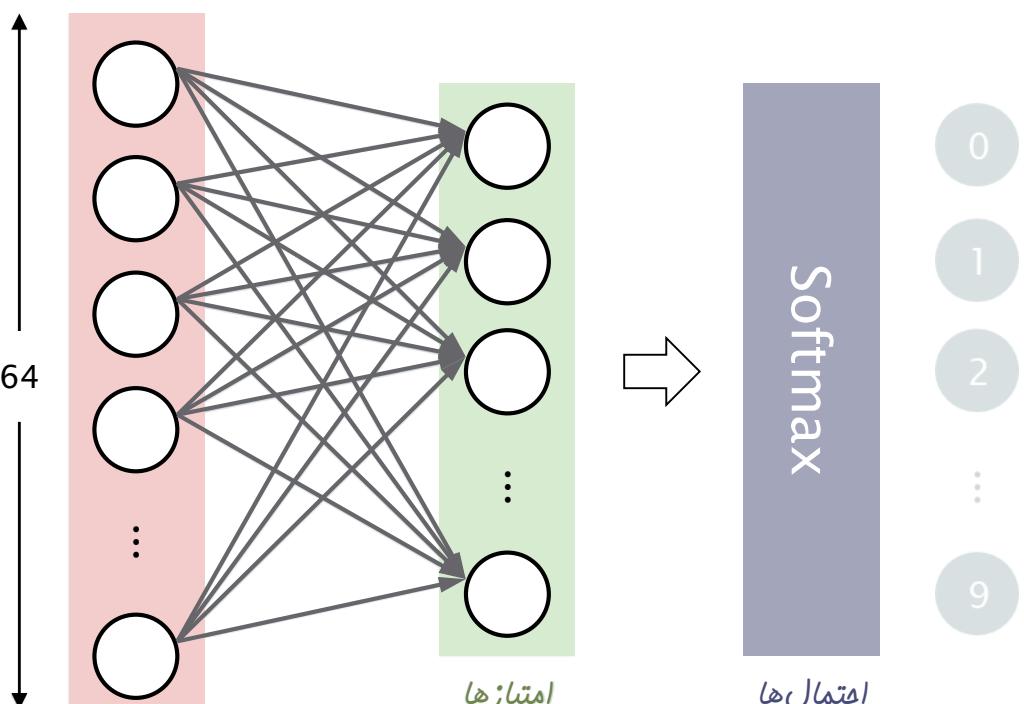
$$p = \text{sigmoid}(x @ w + b)$$

توجه. در صورت استفاده از تابع سیگموید (لजستیک) در یک مسئله دسته‌بندی چند دسته‌ای، دیگر مجموع مقادیر خرضیه‌ها نزوماً برابر با یک نفوادرد بود.

رگرسیون لجستیک چند دسته‌ای: تابع سافت‌مکس

۱۷

□ **تابع سافت‌مکس.** در دسته‌بندی چند دسته‌ای به منظور محاسبه احتمال تعلق بردار ورودی به هر یک از دسته‌های مختلف، به جای تابع سیگموید از تابع سافت‌مکس استفاده می‌کنیم.



$$p = \text{softmax}(x @ w + b)$$

$$\text{softmax}(s^{(i)})_k = \frac{e^{s_k}}{\sum_{j=1}^c e^{s_j}} = p_k^{(i)}$$

احتمال تعلق داده ورودی $x^{(i)}$ به کلاس k

۱۹/۶۵

دسته‌بند سافت‌مکس (رگرسیون لجستیک پنداشتهای)

۱۸

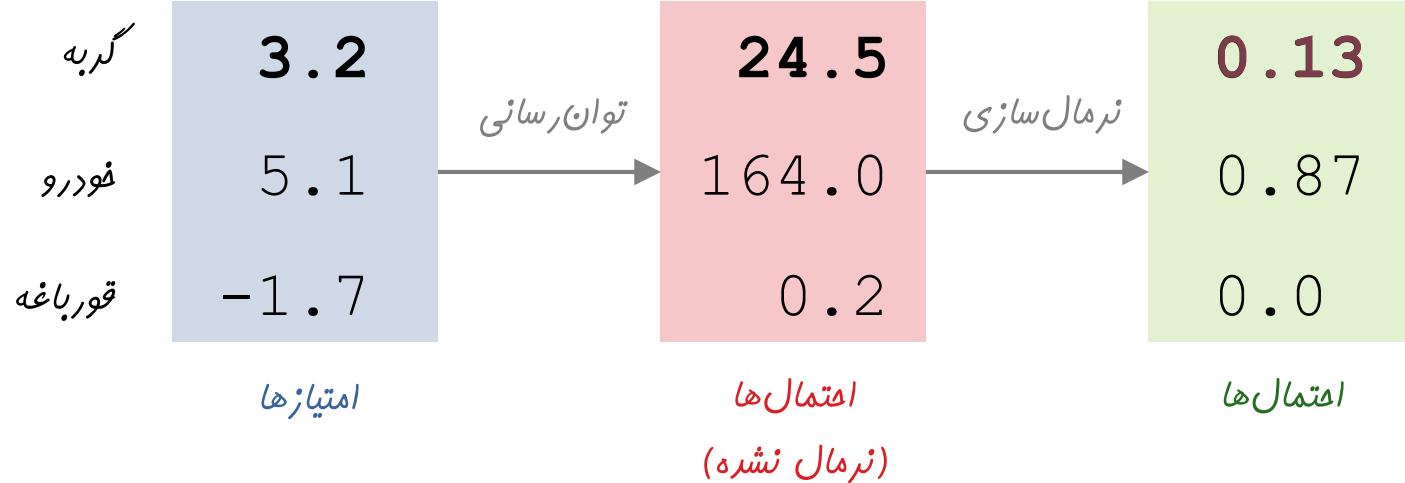
ایده. تبدیل بردار امتیازها به یک بردار توزیع احتمال! □

$$(x^{(i)}, y^{(i)})$$



$$L_i = -\log \left(\frac{e^{s_{y^{(i)}}}}{\sum_j e^{s_j}} \right)$$

$$L_i = -\log(0.13) = 0.89$$



کمترین و بیشترین مقدار ممکن
برای تابع هزینه چقدر است؟

دسته‌بند سافت‌مکس (رگرسیون لجستیک پنداشتهای)

۱۹

$$(x^{(i)}, y^{(i)})$$



□ امتیازها. لگاریتم احتمال دسته‌ها به صورت نرمال نشده!

$$P(Y = k | X = x^{(i)}) = \frac{e^{s_k}}{\sum_j e^{s_j}}$$

$$s = f(x^{(i)}; W)$$

تابع سافت‌مکس

هدف.

گربه	3.2
فودرو	5.1
خورباغه	-1.7

□ بیشینه‌سازی لگاریتم درست‌نمایی (یا کمینه‌سازی منفی لگاریتم درست‌نمایی)!

$$L_i = -\log P(Y = y^{(i)} | X = x^{(i)}) = -\log \left(\frac{e^{s_{y^{(i)}}}}{\sum_j e^{s_j}} \right)$$

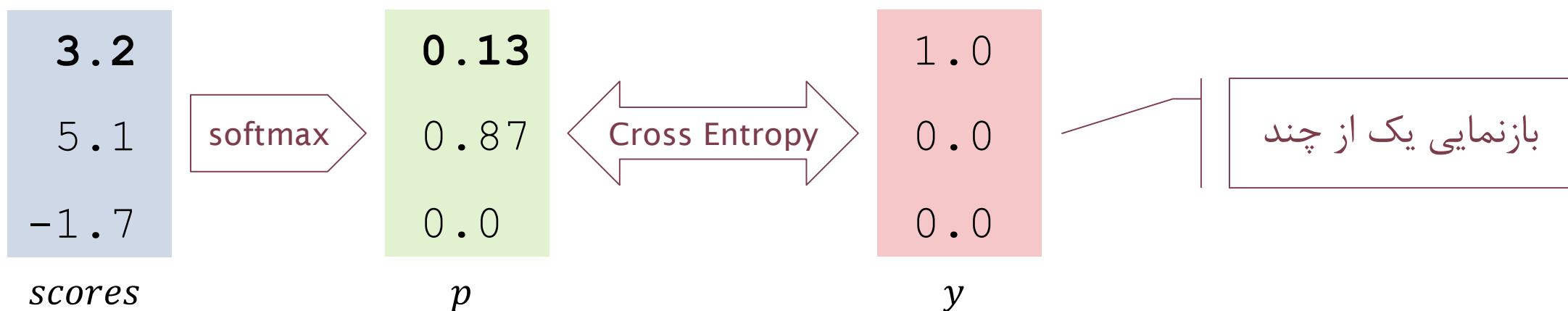
تابع هزینه سافت‌مکس: پیاده‌سازی

۲۰

```
def softmax_loss(scores, y):  
    # softmax loss implementation.  
    # y is encoded as a one-hot vector.  
    p = softmax(scores)  
    return -np.sum(y * np.log(p))
```

$(x^{(i)}, y^{(i)})$

$$L_i = \sum_{k=1}^c -y_k \log p_k = -\log p_{y^{(i)}}$$



دسته‌بند سافت‌مگس: آموزش

۲۱

□ مجموعه آموزشی.

$$\{(x^{(i)}, y^{(i)})\}_{i=1}^m \quad x^{(i)} = [x_1^{(i)} \quad x_2^{(i)} \quad \dots \quad x_n^{(i)}]^T \quad y^{(i)} \in \{1, 2, \dots, c\}$$

□ پارامترها.

$$W \in \mathbb{R}^{n \times c} \quad b \in \mathbb{R}^c$$

□ تابع هزینه.

$$L(W, b) = \frac{1}{m} \sum_{i=1}^m L_i + \lambda R(W) = \frac{1}{m} \sum_{i=1}^m (-\log p_{y^{(i)}}) + \lambda \|W\|_2^2$$

دسته‌بند سافت‌مکس: آموزش

۲۲

$$L(W, b) = \frac{1}{m} \sum_{i=1}^m \left(-\log p_{y^{(i)}} \right) + \lambda \|W\|_2^2$$

تابع هزینه. □

$$= \frac{1}{m} \sum_{i=1}^m \left(-\log \left(\frac{e^{s_{y^{(i)}}}}{\sum_{j=1}^c e^{s_j}} \right) \right) + \lambda \|W\|_2^2$$

$$= \frac{1}{m} \sum_{i=1}^m \left(-\log \left(\frac{e^{\left((w^{y^{(i)}})^T x^{(i)} + b^{y^{(i)}} \right)}}{\sum_{j=1}^c e^{\left((w^{(j)})^T x^{(i)} + b^{(j)} \right)}} \right) \right) + \lambda \|W\|_2^2$$

دسته‌بند سافت‌مگس: آموزش

۲۳

$$L_i = -\log p_{y^{(i)}} \quad p_k = \frac{e^{s_k}}{\sum_{j=1}^c e^{s_j}} \quad s_k = (w^{(k)})^T x^{(i)} + b^{(k)}$$

$$\frac{\partial L_i}{\partial w^{(k)}} = ? \quad k \in \{1, 2, \dots, c\}$$

$$\begin{aligned} \frac{\partial L_i}{\partial w^{(k)}} &= \frac{\partial L_i}{\partial p_{y^{(i)}}} \cdot \frac{\partial p_{y^{(i)}}}{\partial s_k} \cdot \frac{\partial s_k}{\partial w^{(k)}} \\ &= \left(-\frac{1}{p_{y^{(i)}}} \right) p_k (1 - p_k) x^{(i)} \\ &= (p_k - 1) x^{(i)} \end{aligned}$$

$k = y^{(i)}$

$$\begin{aligned} \frac{\partial L_i}{\partial w^{(k)}} &= \frac{\partial L_i}{\partial p_{y^{(i)}}} \cdot \frac{\partial p_{y^{(i)}}}{\partial s_k} \cdot \frac{\partial s_k}{\partial w^{(k)}} \\ &= \left(-\frac{1}{p_{y^{(i)}}} \right) p_{y^{(i)}} (-p_k) x^{(i)} \\ &= (p_k) x^{(i)} \end{aligned}$$

$k \neq y^{(i)}$

دسته‌بند سافت‌مگس: آموزش

۲۴

$$L_i = -\log p_{y^{(i)}} = -\log \left(\frac{e^{s_{y^{(i)}}}}{\sum_{j=1}^c e^{s_j}} \right)$$

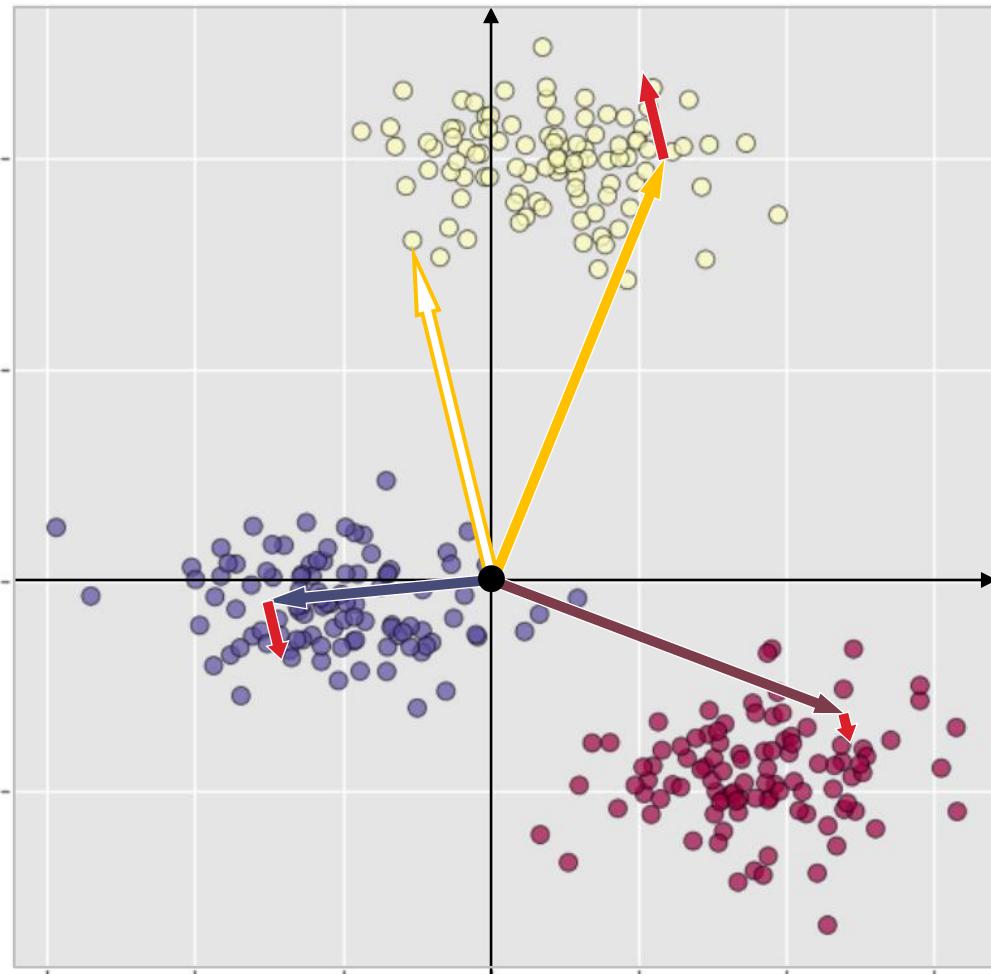
□ الگوریتم گرادیان کاہشی اتفاقی.

$$\frac{\partial L_i}{\partial w^{(k)}} = (p_{y_k} - 1)x^{(i)} \rightarrow w^{(k)} = w^{(k)} - \alpha(p_{y_k} - 1)x^{(i)} \quad (k = y^{(i)})$$

$$\frac{\partial L_i}{\partial w^{(k)}} = (p_{y_k})x^{(i)} \rightarrow w^{(k)} = w^{(k)} - \alpha(p_{y_k})x^{(i)} \quad (k \neq y^{(i)})$$

دسته‌بند سافت‌مگس: آموزش

۲۵



تفسیر هندسی. □

$$w^{(k)} = w^{(k)} - \alpha(p_{y_k} - 1)x^{(i)} \quad (k = y^{(i)})$$

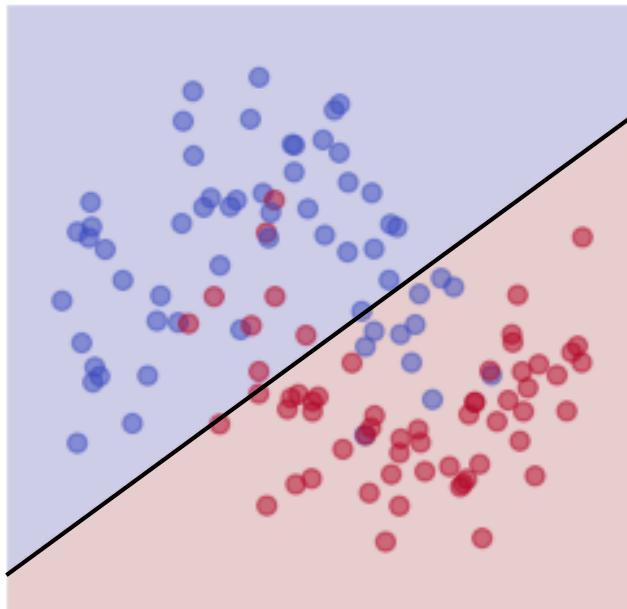
$$w^{(k)} = w^{(k)} - \alpha p_{y_k} x^{(i)} \quad (k \neq y^{(i)})$$

شبکه‌های عمیقی

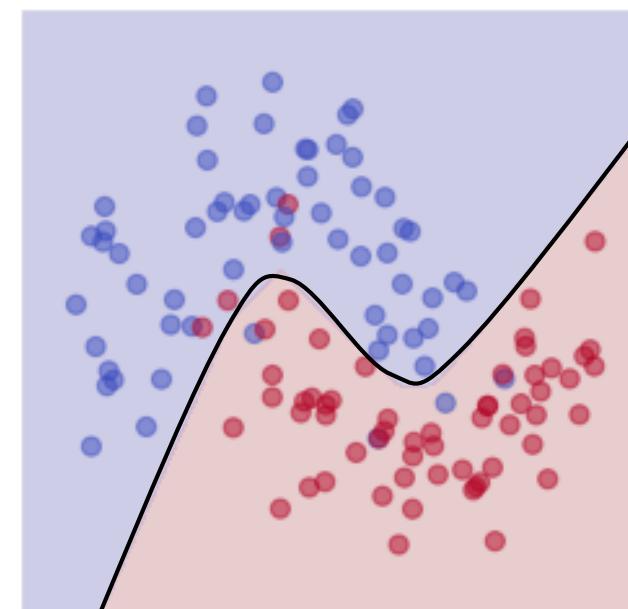
شبکه‌های عصبی: انگیزه

۲۷

- رگرسیون لجستیک یک روش دسته‌بندی خطی است.
- اگر داده‌ها به صورت خطی تفکیک‌پذیر نباشند، نیاز به افزودن ویژگی‌های مرتبه بالاتر داریم.



مرز تصمیم‌گیری خطی



مرز تصمیم‌گیری غیرخطی

شبکه‌های عصبی: انگیزه

۲۸

- رگرسیون لجستیک یک روش دسته‌بندی خطی است.
- اگر داده‌ها به صورت خطی تفکیک‌پذیر نباشند، نیاز به افزودن ویژگی‌های مرتبه بالاتر داریم.

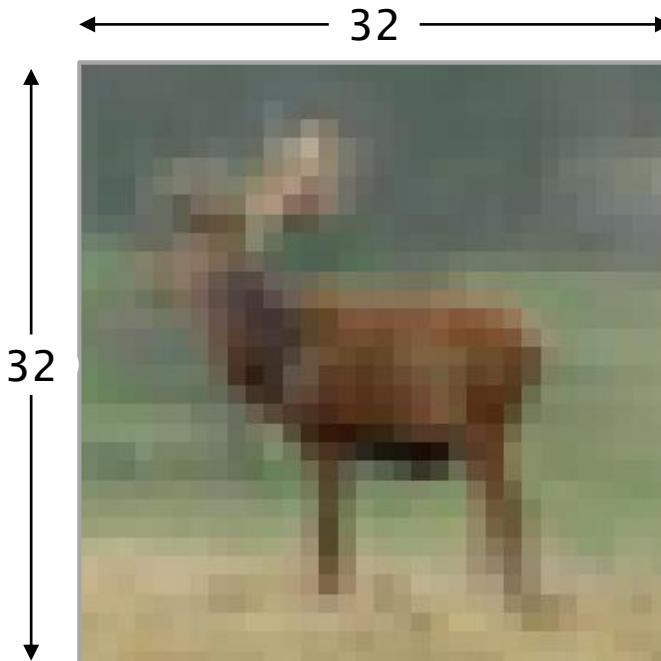
8								
8	0	0	13	16	16	15	2	0
	0	0	14	13	11	16	2	0
	0	0	11	13	15	6	0	0
	0	0	5	16	10	0	0	0
	0	0	10	14	15	0	0	0
	0	1	14	3	15	7	0	0
	0	6	11	0	15	6	0	0
	0	1	13	16	15	3	0	0

- مثال. تشخیص ارقام دستنویس [تصاویر ۸ در ۸]
- تعداد ویژگی‌های مرتبه دوم: بیش از ۲,۰۰۰
- تعداد ویژگی‌های مرتبه سوم: بیش از ۴۰,۰۰۰
- ... □

شبکه‌های عصبی: انگیزه

۲۹

- رگرسیون لجستیک یک روش دسته‌بندی خطی است.
- اگر داده‌ها به صورت خطی تفکیک‌پذیر نباشند، نیاز به افزودن ویژگی‌های مرتبه بالاتر داریم.



□ مثال. تصاویر رنگی [۳۲ در ۳۲ در ۳]

□ تعداد ویژگی‌های مرتبه دوم: بیش از ۴۷۰,۰۰۰

□ تعداد ویژگی‌های مرتبه سوم: بیش از ۵,۰۰۰,۰۰۰

... □

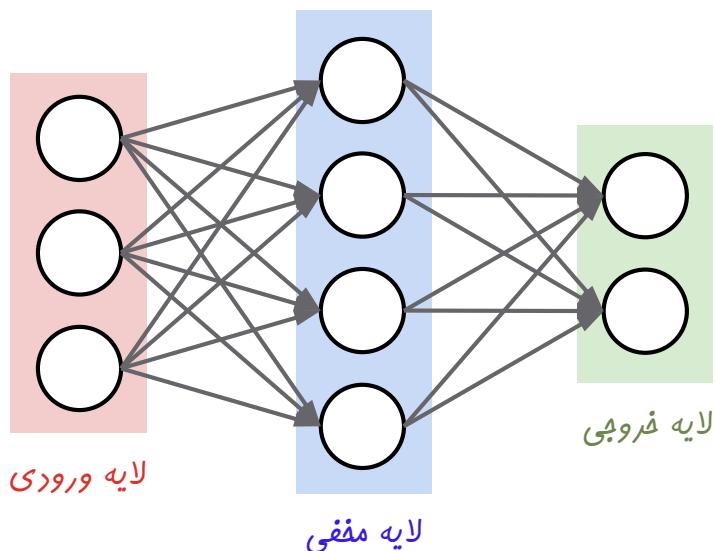


می‌توانیم بسیاری از ویژگی‌ها را با استفاده از تنظیم حذف کنیم!!!

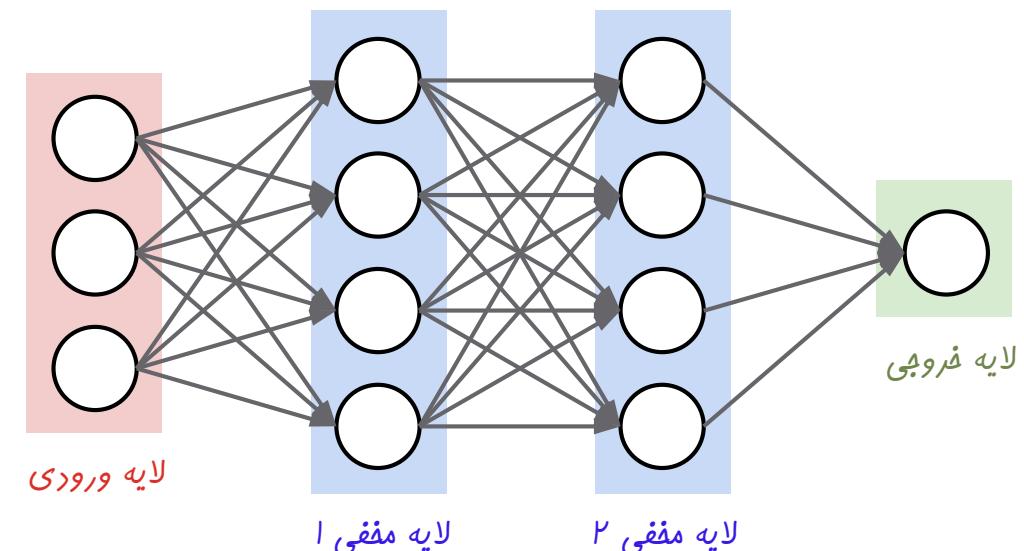
شبکه‌های عصبی: یادگیری ویژگی‌های جدید

۳۰

- شبکه‌های عصبی می‌توانند با ترکیب ویژگی‌های سطح پایین، ویژگی‌های سطح بالای مورد نیاز خود را یاد بگیرند.



شبکه عصبی ۲ لایه

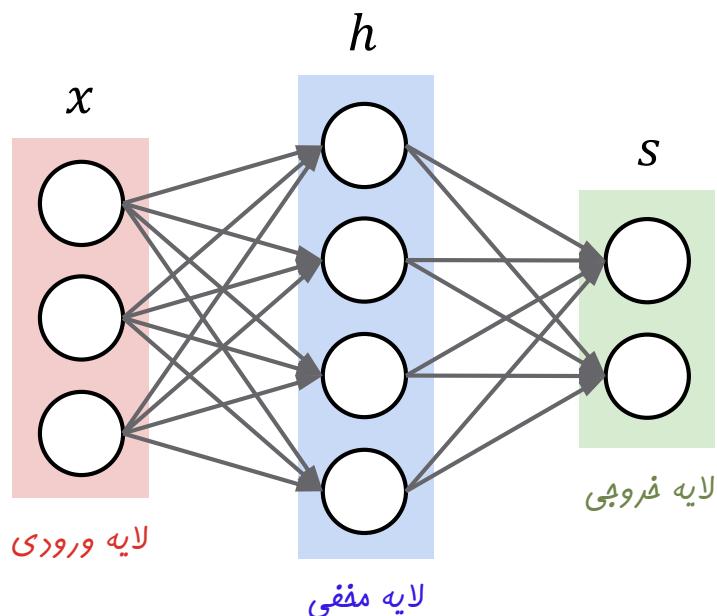


شبکه عصبی ۳ لایه

شبکه‌های عصبی: یادگیری ویژگی‌های جدید

۲۱

□ شبکه‌های عصبی می‌توانند با ترکیب ویژگی‌های سطح پایین، ویژگی‌های سطح بالای مورد نیاز خود را یاد بگیرند.



$$s = Wx + b$$

دسته‌بندی خطی

$$h = f(W_1x + b_1)$$

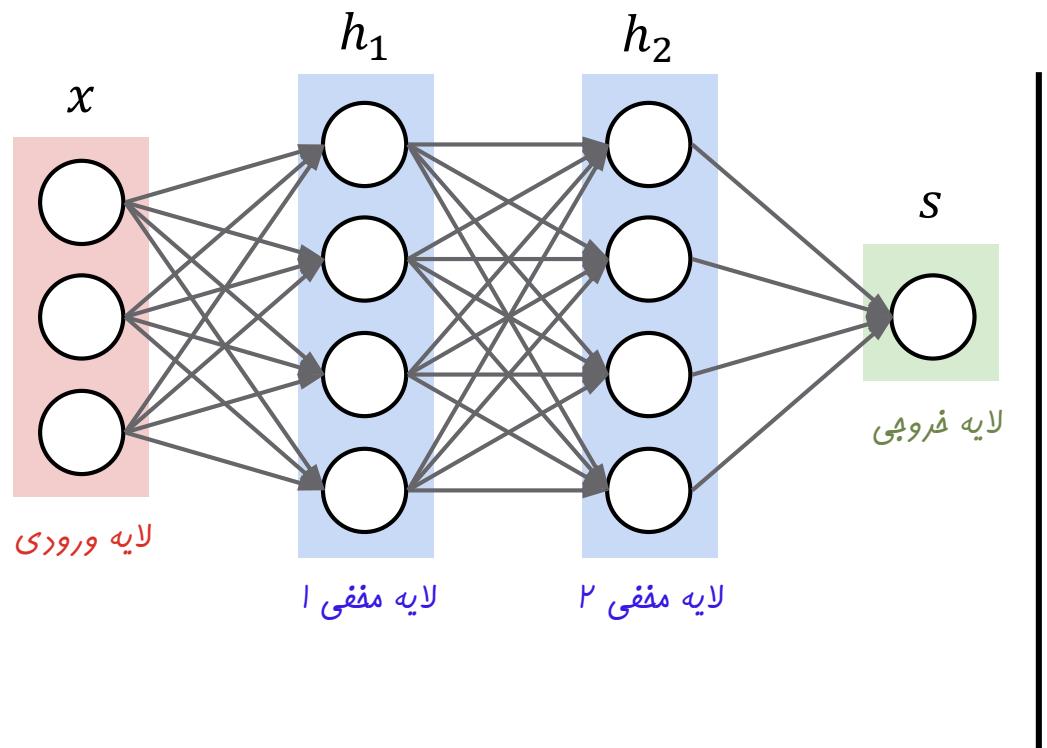
شبکه عصبی دو لایه

$$s = W_2h + b_2$$

شبکه‌های عصبی: یادگیری ویژگی‌های جدید

۳۲

شبکه‌های عصبی می‌توانند با ترکیب ویژگی‌های سطح پایین، ویژگی‌های سطح بالای مورد نیاز خود را یاد بگیرند. □



$$h = f(W_1x + b_1)$$

شبکه عصبی دو لایه

$$s = W_2h + b_2$$

$$h_1 = f(W_1x + b_1)$$

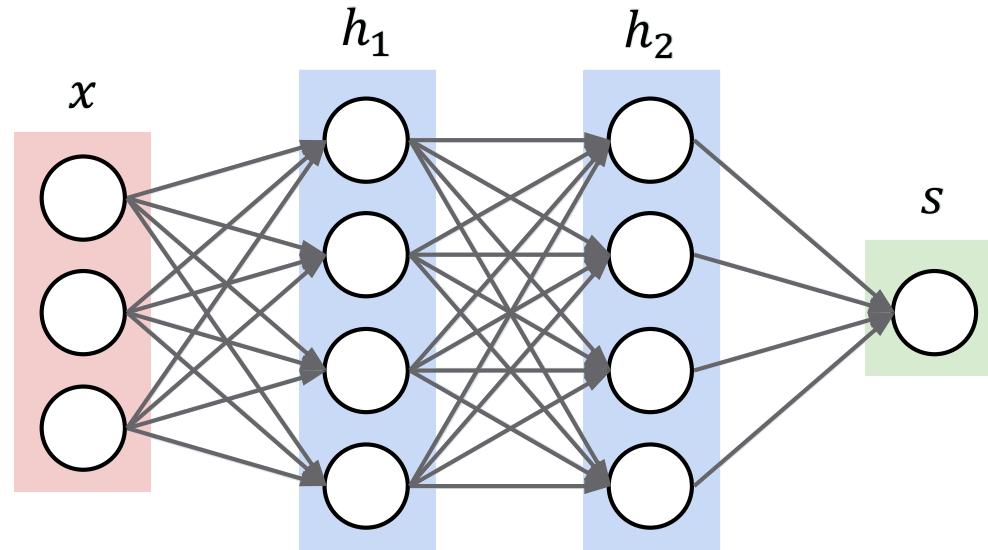
شبکه عصبی سه لایه

$$h_2 = f(W_2h_1 + b_2)$$

$$s = W_3h_2 + b_3$$

شبکه‌های عصبی: پیاده‌سازی انتشار پیش‌(و)

۳۲



```
f = lambda x: 1.0 / (1.0 + np.exp(-x))      # activation function (sigmoid)

x = np.random.randn(3, 1)                      # random input vector (3x1)

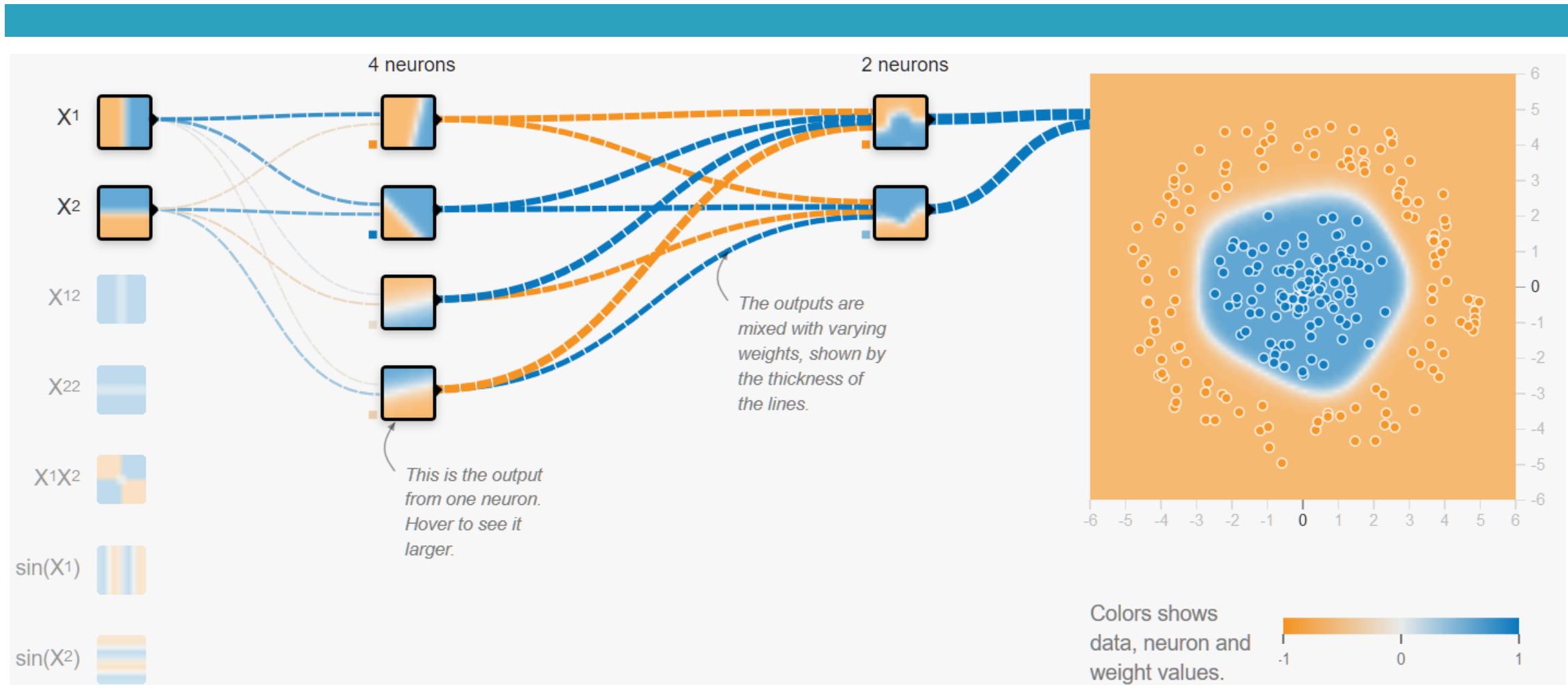
h1 = f(W1 @ x + b1)                          # first hidden layer activations (4x1)

h2 = f(W2 @ h1 + b2)                          # second hidden layer activations (4x1)

s = W3 @ h2 + b3                                # scores (1x1)
```

اجرای نمایشی

۲۴



<https://playground.tensorflow.org>

توابع فعالیت

۳۵

□ شبکه عصبی سه لایه.

$$s = W_3 f(W_2 f(W_1 x))$$

□ اهمیت توابع فعالیت غیرخطی در لایه‌های مخفی.

□ عدم استفاده از توابع فعالیت غیرخطی در لایه‌های مخفی، باعث می‌شود شبکه عصبی به یک **دسته‌بند خطی** ساده تبدیل گردد!

$$s = W_3(W_2(W_1 x))$$

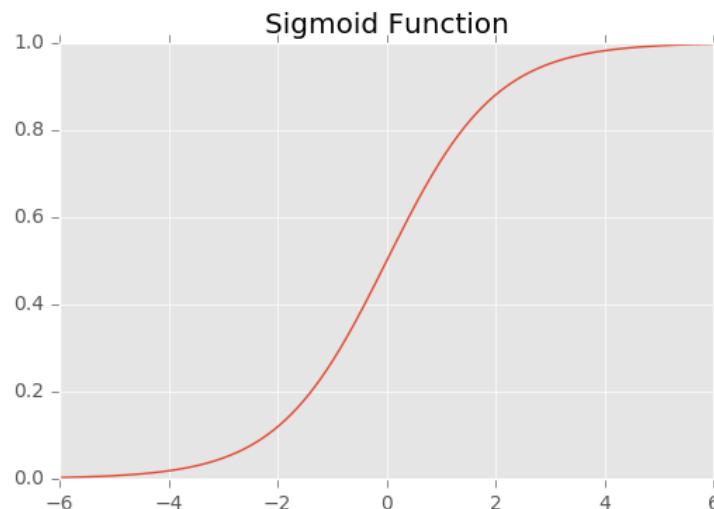
$$= (W_3 W_2 W_1)x$$

$$= Wx$$

توابع فعالیت

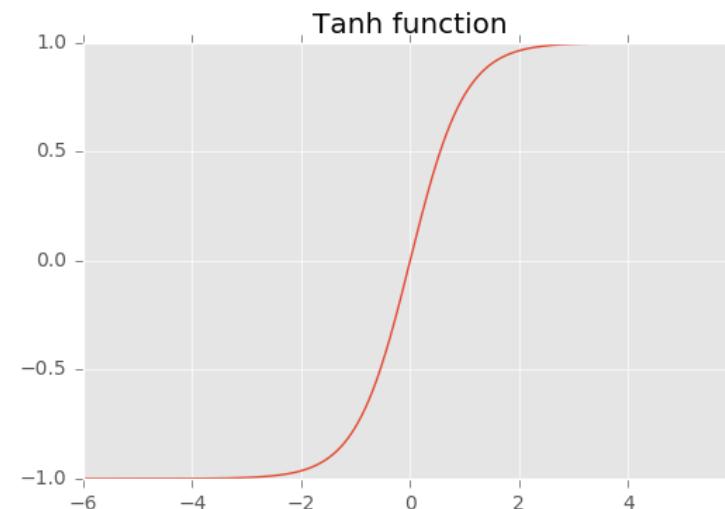
۳۶

سیگموید



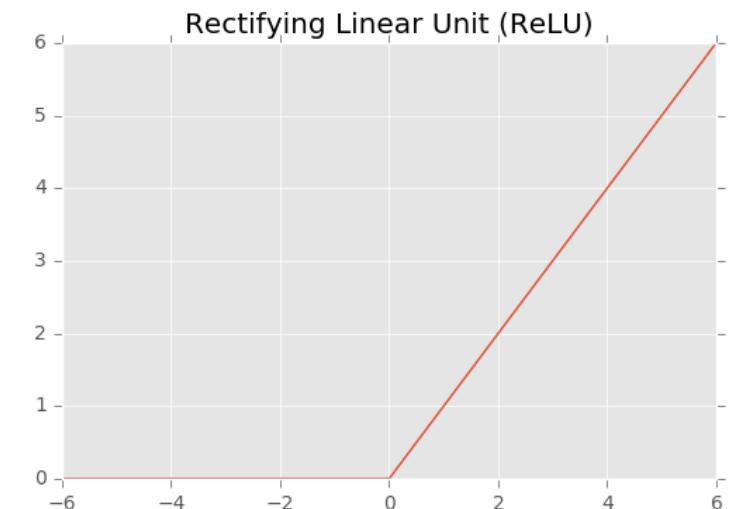
$$\sigma(x) = 1/(1 + e^{-x})$$

تانژانت هایپربولیک



$$\tanh(x)$$

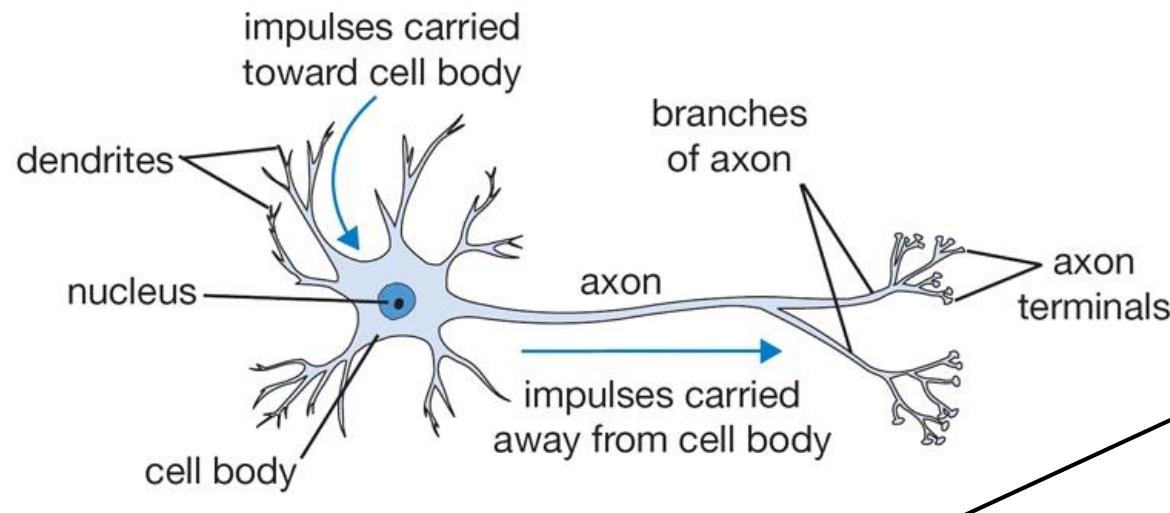
ReLU



$$\max(0, x)$$

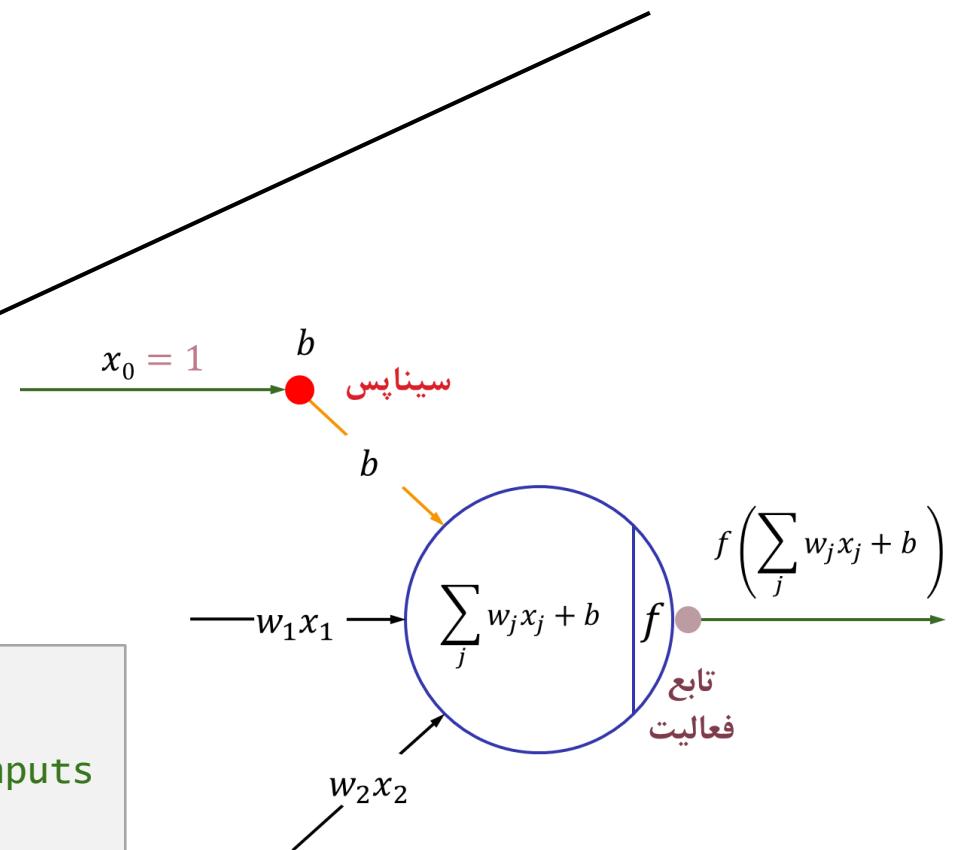
نورون‌ها و شبکه‌های عصبی

۳۷



```
# assume w and x are 1d numpy arrays
```

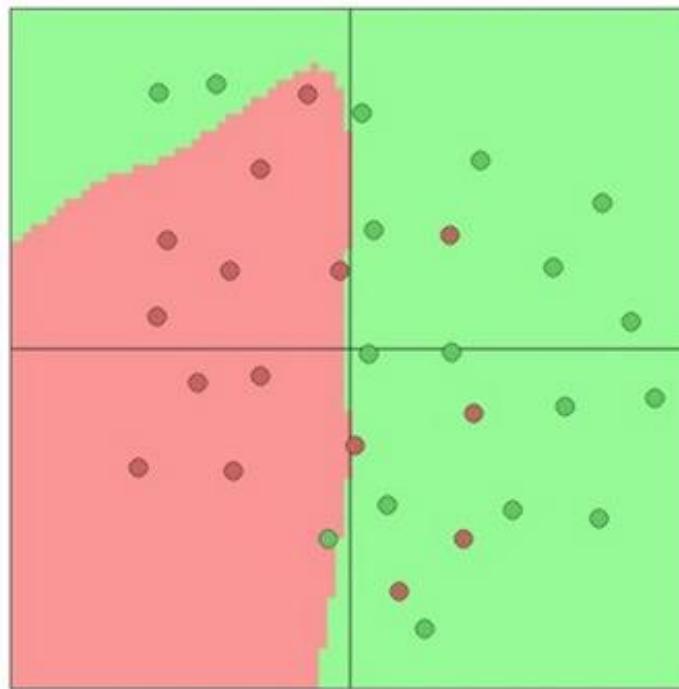
```
net_input = np.sum(w * x) + b          # weighted sum of inputs  
output = 1.0 / (1.0 + np.exp(-net_input)) # sigmoid function
```



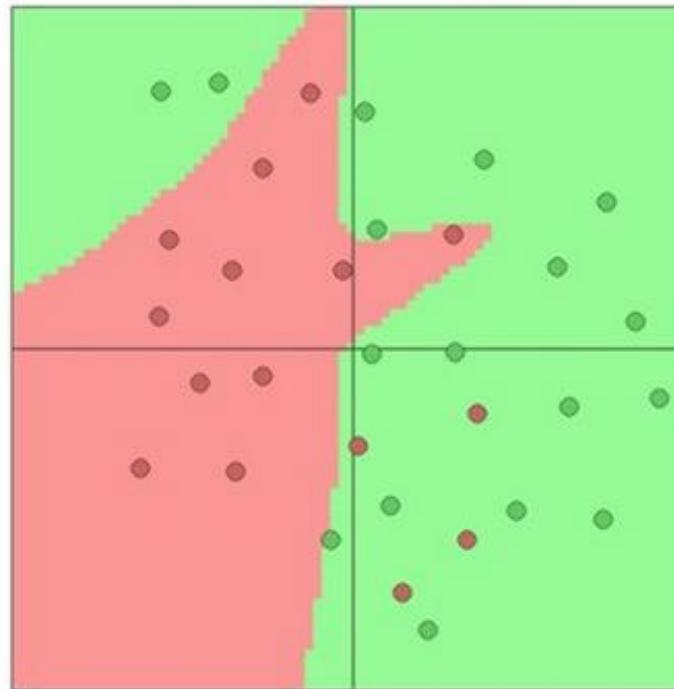
تعیین تعداد و اندازه لایه‌ها

۳۸

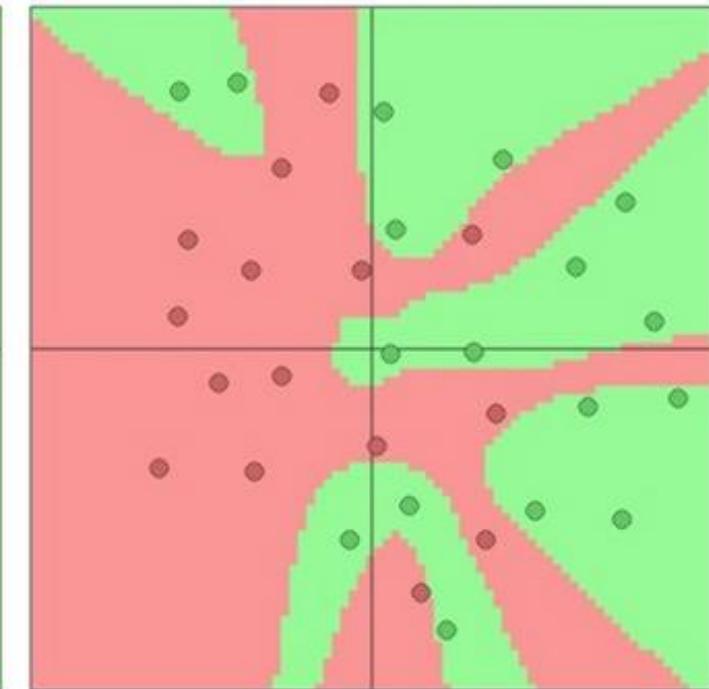
۳ نورون در لایه مخفی



۶ نورون در لایه مخفی



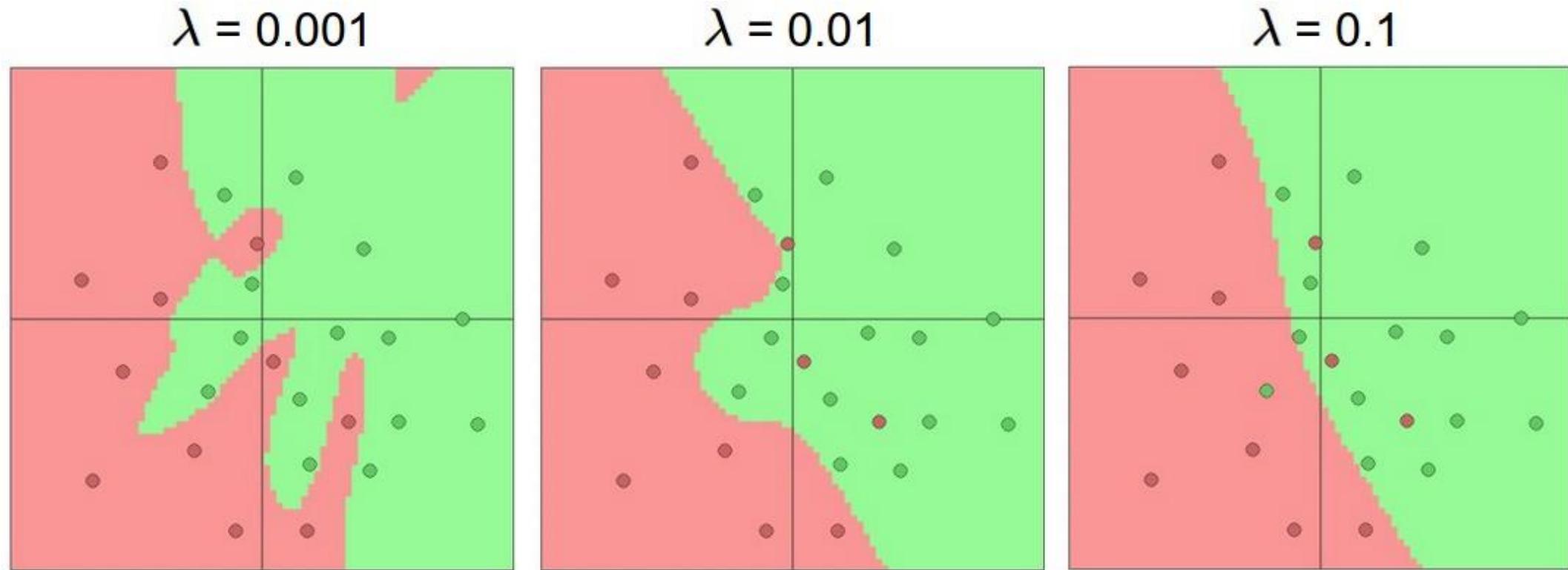
۲۰ نورون در لایه مخفی



نورون‌های بیشتر = ظرفیت بیشتر

تعیین تعداد و اندازه لایه‌ها

۳۹



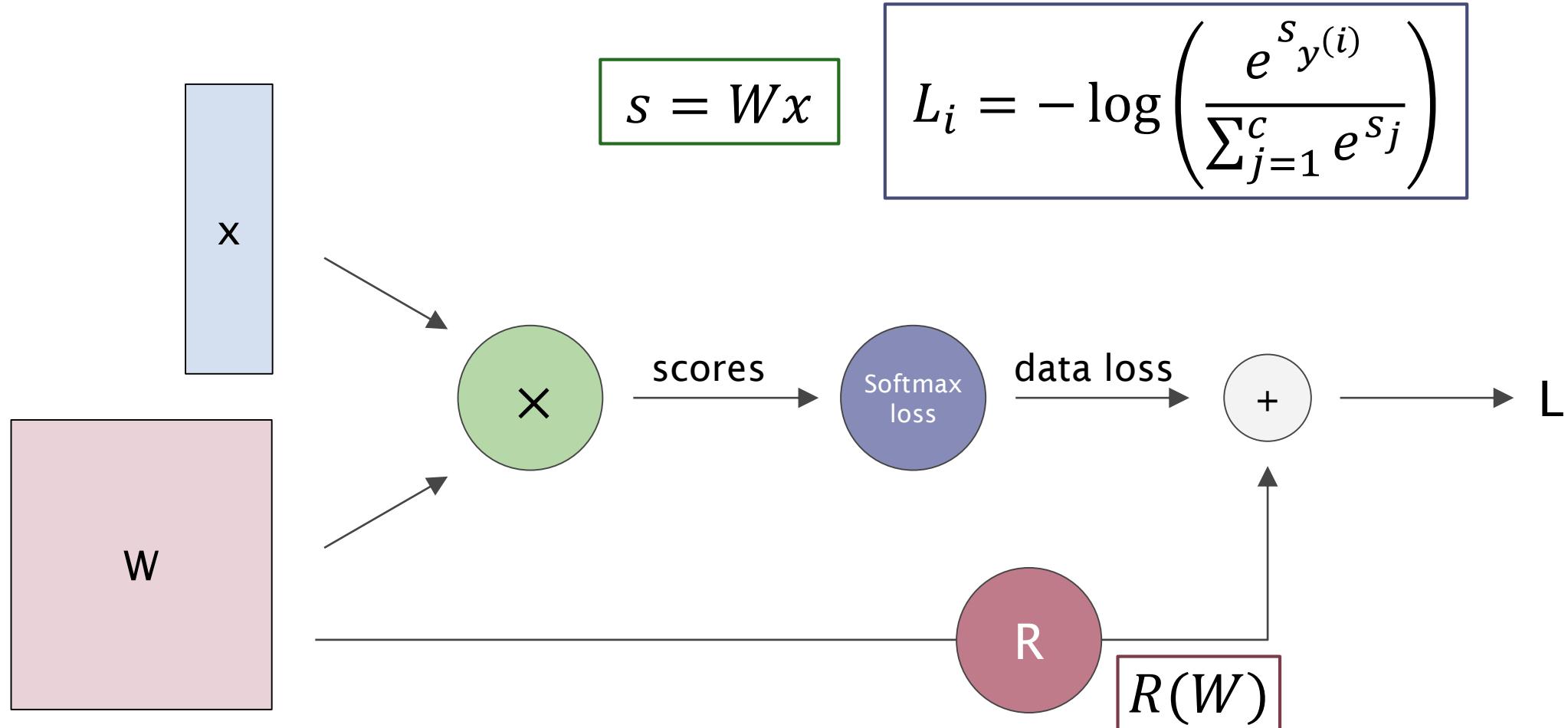
از اندازه شبکه عصبی به عنوان تنظیم کننده استفاده نکنید و به جای آن از یک روش قوی‌تر استفاده کنید. □

L2
تنظیم

الگوئیم پس انتشار خطا

دسته‌بند سافت‌مکس: گراف دماسپاری

۴۱

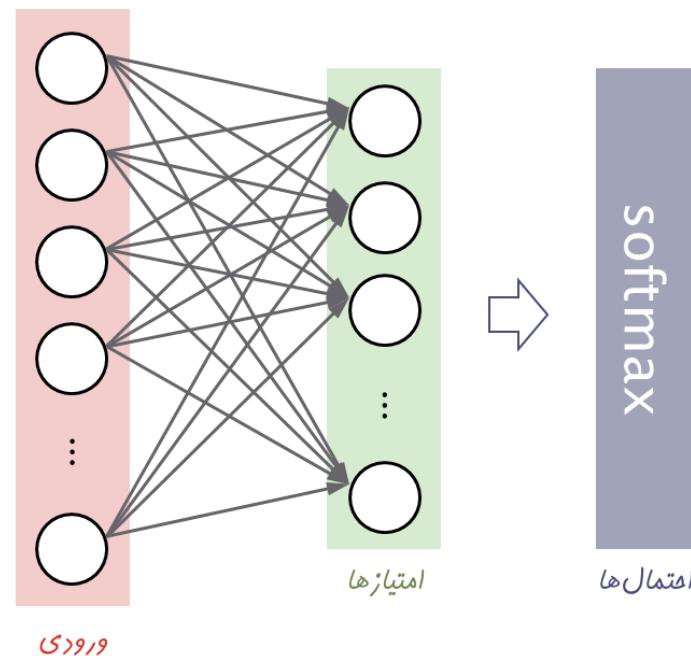


یادآوری: محاسبه گرادیان‌ها در دسته‌بند سافت‌مکس

۴۲

$$L = \frac{1}{m} \left(\sum_{i=1}^m -\log \left(\frac{e^{s_{y(i)}}}{\sum_{j=1}^c e^{s_j}} \right) \right) + \lambda \|W\|_2^2$$

تابع هزینه. □



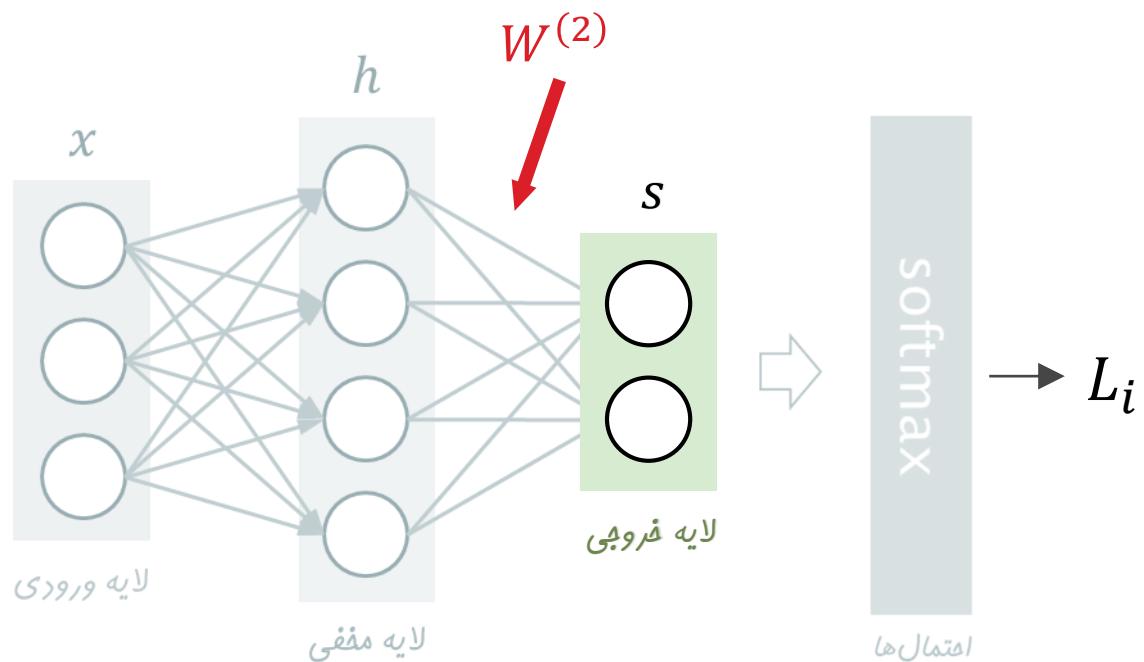
محاسبه گرادیان‌ها. □

$$\begin{aligned}\frac{\partial L_i}{\partial w^{(k)}} &= \frac{\partial L_i}{\partial s_k} \cdot \frac{\partial s_k}{\partial w^{(k)}} \\ &= (p_k - \{y^{(i)} == k\}) x^{(i)}\end{aligned}$$

$$\{y^{(i)} == k\} = \begin{cases} 1, & y^{(i)} = k \\ 0, & y^{(i)} \neq k \end{cases}$$

الگوریتم پس انتشار: مماسه گرادیانها

۴۳



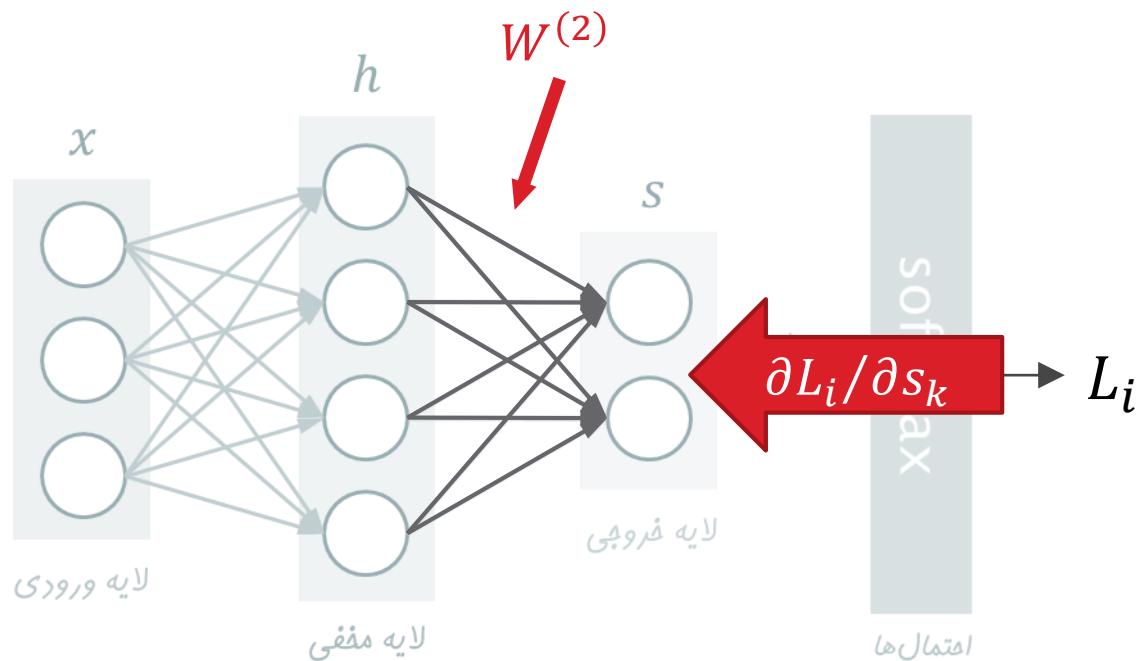
$$\frac{\partial L_i}{\partial W^{(2)}} = \frac{\partial L_i}{\partial s_k} \cdot \frac{\partial s_k}{\partial W^{(2)}}$$

$$= (p_k - \{y^{(i)} == k\})$$

$$L = \frac{1}{m} \left(\sum_{i=1}^m -\log \left(\frac{e^{s_{y^{(i)}}}}{\sum_{j=1}^c e^{s_j}} \right) \right) + \lambda \|W\|_2^2$$

الگوریتم پس انتشار: مهاسبه گرادیان‌ها

۴۴



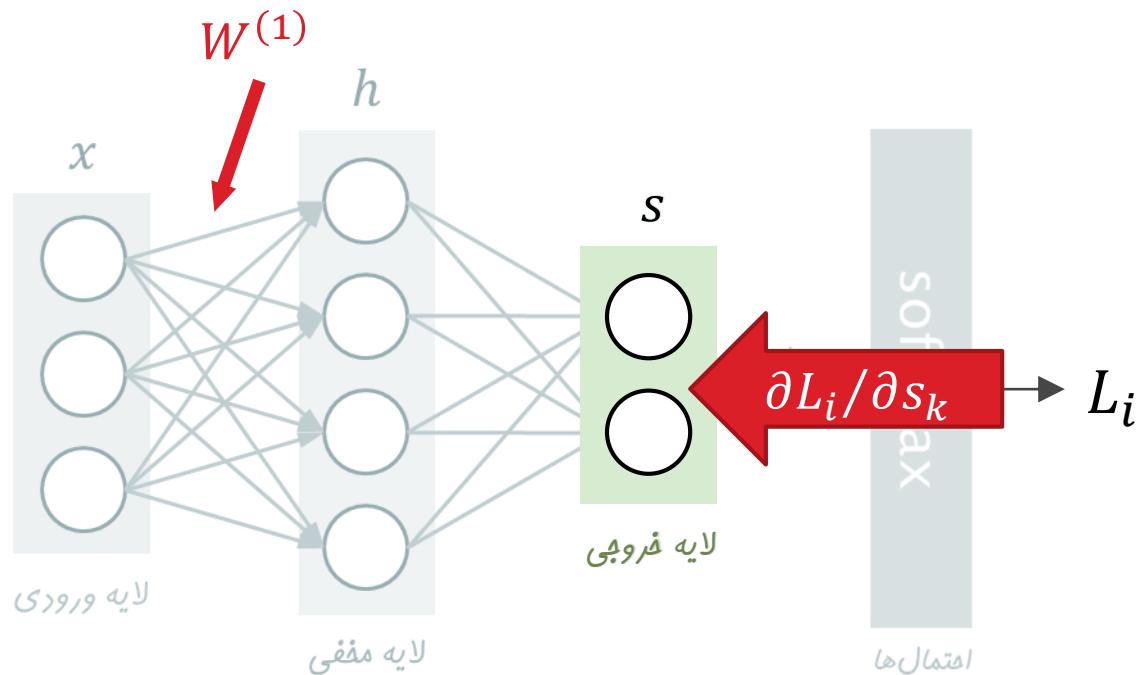
$$\frac{\partial L_i}{\partial W^{(2)}} = \frac{\partial L_i}{\partial s_k} \cdot \frac{\partial s_k}{\partial W^{(2)}}$$

$$= (p_k - \{y^{(i)} == k\}) \cdot h$$

$$L = \frac{1}{m} \left(\sum_{i=1}^m -\log \left(\frac{e^{s_{y^{(i)}}}}{\sum_{j=1}^c e^{s_j}} \right) \right) + \lambda \|W\|_2^2$$

الگوریتم پس انتشار: مماسه گرادیانها

۴۵

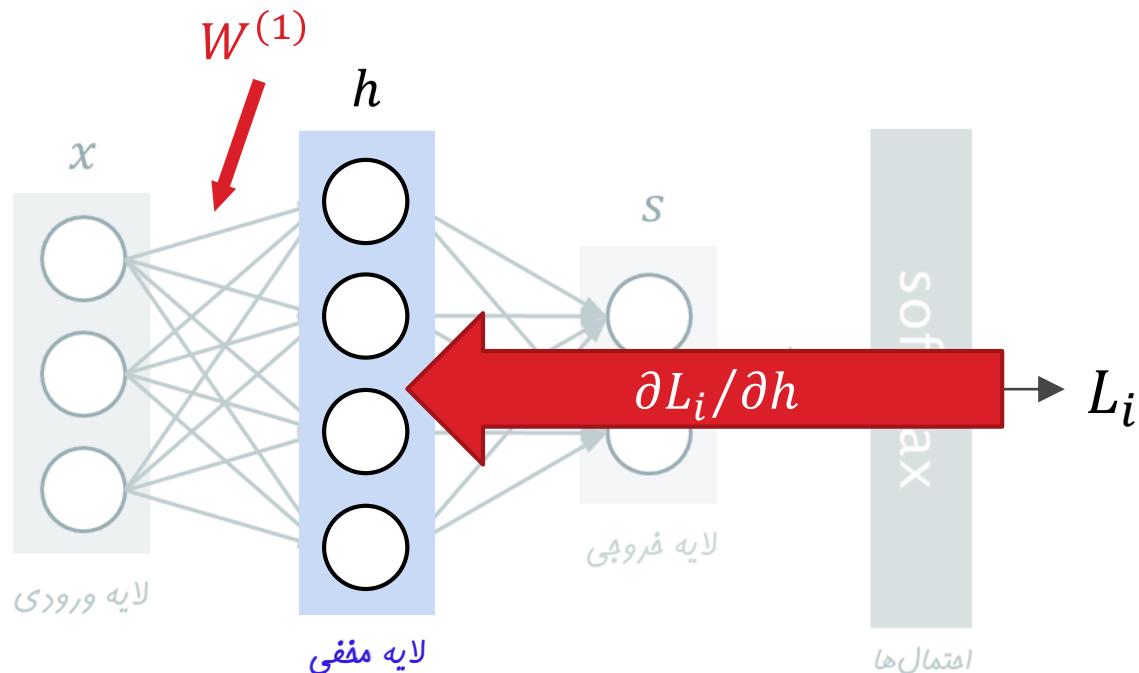


$$\frac{\partial L_i}{\partial W^{(1)}} = \frac{\partial L_i}{\partial s_k} \cdot \frac{\partial s_k}{\partial h} \cdot \frac{\partial h}{\partial f} \cdot \frac{\partial f}{\partial W^{(1)}}$$

$$L = \frac{1}{m} \left(\sum_{i=1}^m -\log \left(\frac{e^{s_{y(i)}}}{\sum_{j=1}^c e^{s_j}} \right) \right) + \lambda \|W\|_2^2$$

الگوریتم پس انتشار: مهاسبه گرادیانها

۴۶

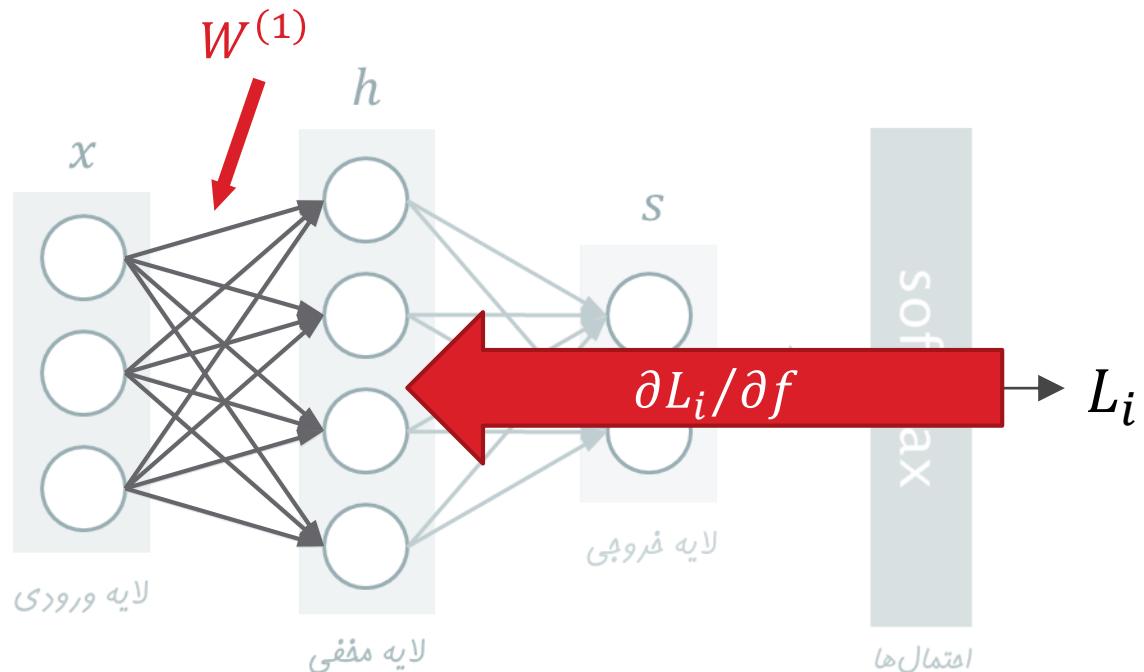


$$\begin{aligned} \frac{\partial L_i}{\partial W^{(1)}} &= \underbrace{\frac{\partial L_i}{\partial s_k}}_{\text{Red}} \cdot \underbrace{\frac{\partial s_k}{\partial h}}_{\text{Blue}} \cdot \frac{\partial h}{\partial f} \cdot \frac{\partial f}{\partial W^{(1)}} \\ &= \frac{\partial L_i}{\partial h} \cdot \frac{\partial h}{\partial f} \cdot \frac{\partial f}{\partial W^{(1)}} \end{aligned}$$

$$L = \frac{1}{m} \left(\sum_{i=1}^m -\log \left(\frac{e^{s_{y(i)}}}{\sum_{j=1}^c e^{s_j}} \right) \right) + \lambda \|W\|_2^2$$

الگوریتم پس انتشار: مهاسبه گرادیانها

۴۷

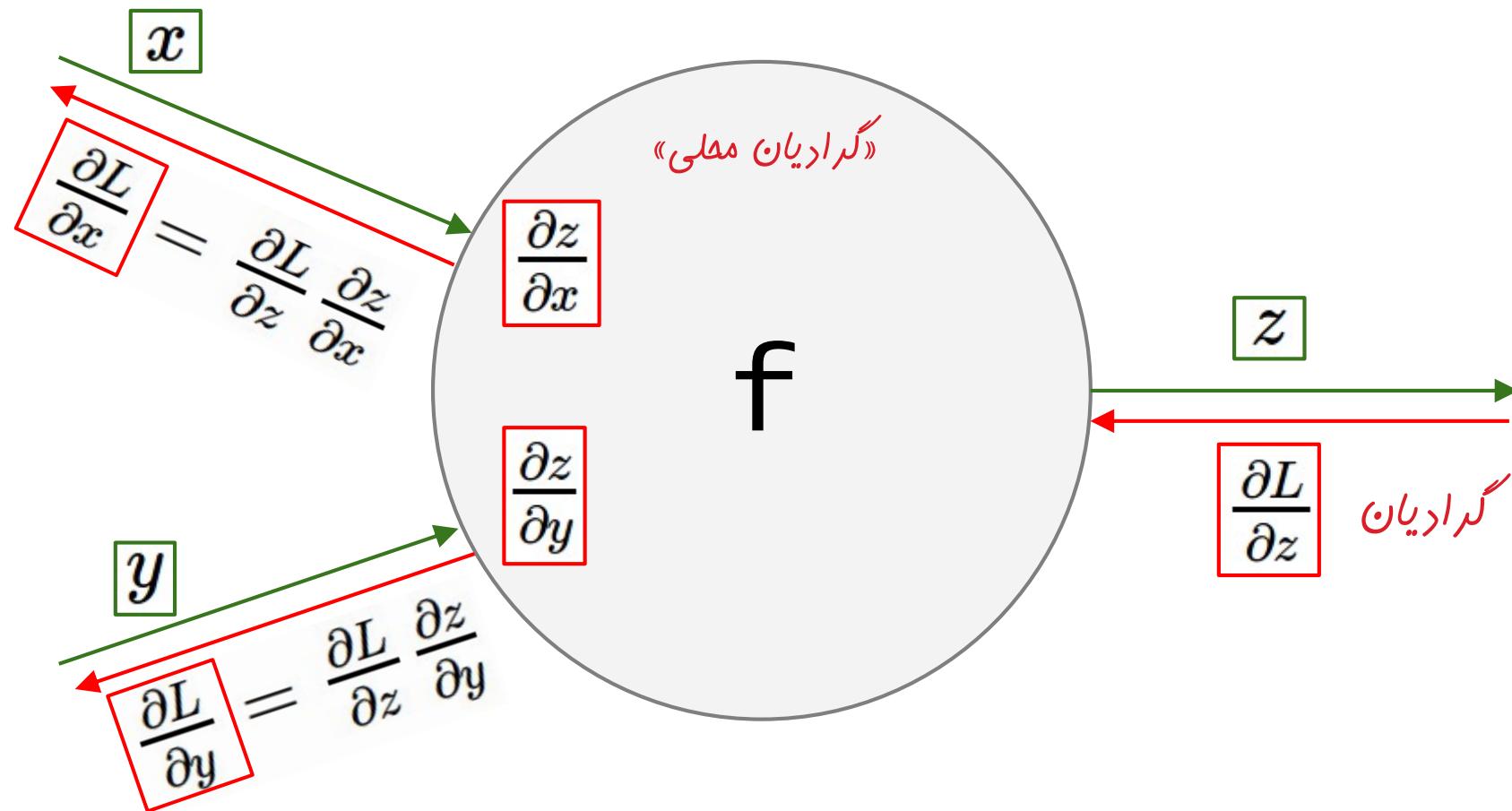


$$\begin{aligned}
 \frac{\partial L_i}{\partial W^{(1)}} &= \underbrace{\frac{\partial L_i}{\partial s_k}}_{\text{Red}} \cdot \underbrace{\frac{\partial s_k}{\partial h}}_{\text{Blue}} \cdot \underbrace{\frac{\partial h}{\partial f}}_{\text{Blue}} \cdot \underbrace{\frac{\partial f}{\partial W^{(1)}}}_{\text{Blue}} \\
 &= \frac{\partial L_i}{\partial h} \cdot \frac{\partial h}{\partial f} \cdot \frac{\partial f}{\partial W^{(1)}} \\
 &= \frac{\partial L_i}{\partial f} \cdot \frac{\partial f}{\partial W^{(1)}}
 \end{aligned}$$

$$L = \frac{1}{m} \left(\sum_{i=1}^m -\log \left(\frac{e^{s_{y(i)}}}{\sum_{j=1}^c e^{s_j}} \right) \right) + \lambda \|W\|_2^2$$

الگوریتم پس انتشار: محاسبه گرادیان‌ها

۴۸



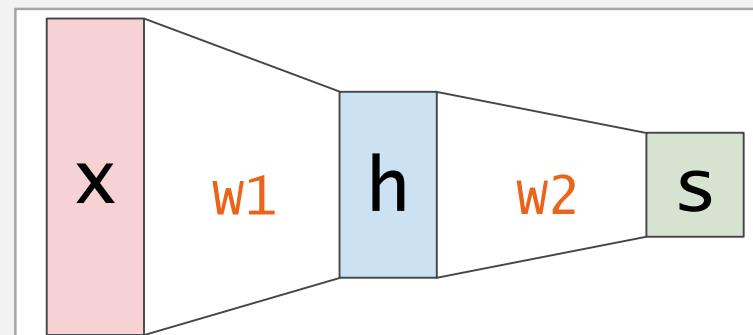
پیادهسازی یک شبکه عصبی دو لایه

۴۹

```
# receive w1, w2, b1, b2 (weights/biases), x (data)

# forward pass:
h =          #... function of x, w1, b1
scores =     #... function of h, w2, b2
loss =       #... (several lines of code to evaluate Softmax loss)

# backward pass:
dscores =    #... dL/dscores
dh, dw2, db2 = #... dL/dh, dL/dw2, dL/db2
dw1, db1 =   #... dL/dw1, dL/db1
```



جمع‌بندی

۵۰

- تعداد پارامترها در یک شبکه عصبی می‌تواند بسیار زیاد باشد:
 - نوشتن رابطه مربوط به گرادیان تمام پارامترها به صورت دستی غیر ممکن است!
- پس انتشار. به کار بردن **قاعده زنجیری** به صورت بازگشته در طول یک گراف محاسباتی به منظور محاسبه گرادیان تابع هزینه نسبت به پارامترها، ورودی‌ها و مقادیر میانی.
- **گراف محاسباتی**. یک ساختار گرافی که هر گره آن محاسبات رو به جلو و محاسبات رو به عقب را پیاده‌سازی می‌کند.
- محاسبات رو به جلو. محاسبه نتیجه یک عمل و ذخیره مقادیر میانی مورد نیاز برای محاسبه گرادیان.
- محاسبات رو به عقب. استفاده از قاعده زنجیری به منظور محاسبه گرادیان تابع هزینه نسبت به ورودی‌ها.

روش‌های بهینه‌سازی پیش‌رفته

بهینه‌سازی پیش‌رفته

۵۲

```
from scipy.optimize import minimize  
  
minimize(J, x0, args=(X_train, y_train), method='CG', jac=True)
```

یادگیری ماشین / شبکه های عصبی

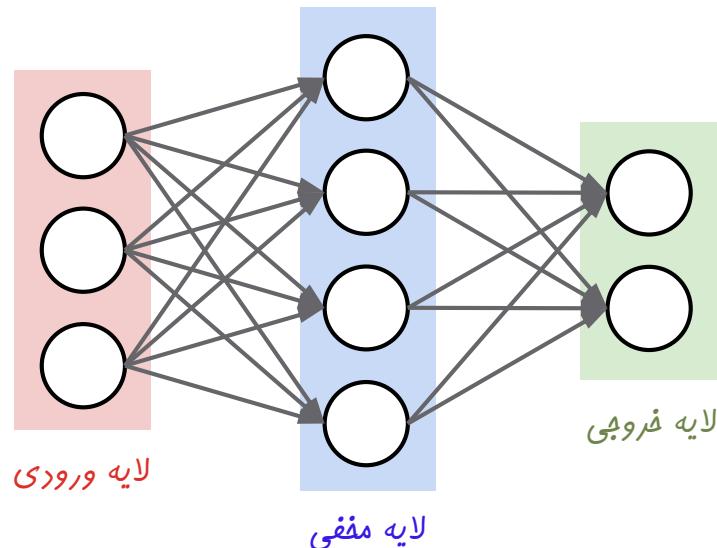
پارامترها و گرادیان‌ها باید به صورت یک بردار باشند

$$L = \frac{1}{m} \left(\sum_{i=1}^m -\log \left(\frac{e^{s_{y(i)}}}{\sum_{j=1}^c e^{s_j}} \right) \right) + \lambda \|W\|_2^2$$

بهینه‌سازی پیش‌رفته

۵۳

□ ارسال پارامترها. پیش از ارسال پارامترها به تابع بهینه‌سازی باید همه آنها را به یک بردار تبدیل کنیم.



$$F = 784 \quad H = 20 \quad C = 10$$

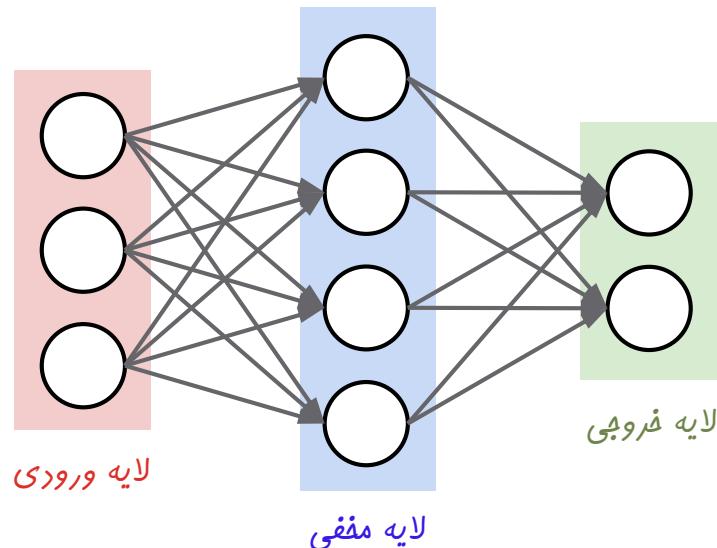
$$\begin{array}{ll} W^{(1)} \in \mathbb{R}^{F \times H} & b^{(1)} \in \mathbb{R}^H \\ W^{(2)} \in \mathbb{R}^{H \times C} & b^{(2)} \in \mathbb{R}^C \end{array} \rightarrow W$$

```
W = np.concatenate((W1.ravel(), b1, W2.ravel(), b2), axis=0)
```

بهینه‌سازی پیش‌رفته

۵۴

دریافت گرادیان‌ها. پیش از دریافت گرادیان‌ها، باید همه آنها را به یک بردار تبدیل کنیم. □



$$F = 784 \quad H = 20 \quad C = 10$$

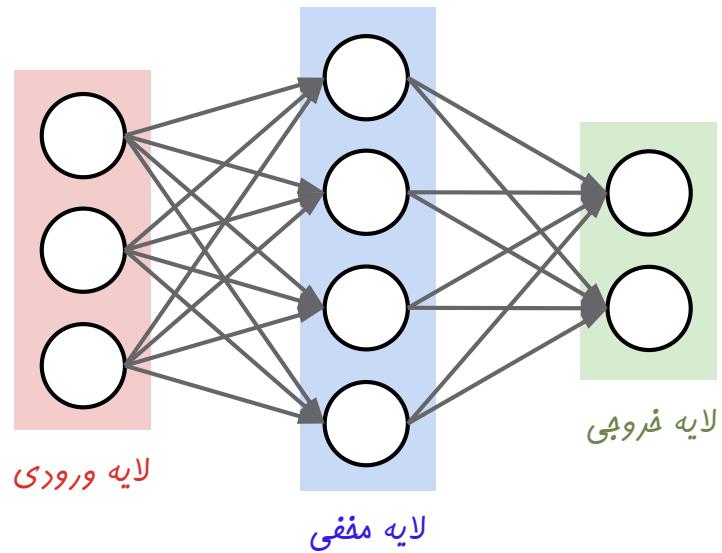
$$\begin{array}{ll} dW^{(1)} \in \mathbb{R}^{F \times H} & db^{(1)} \in \mathbb{R}^H \\ dW^{(2)} \in \mathbb{R}^{H \times C} & db^{(2)} \in \mathbb{R}^C \end{array} \rightarrow dW$$

```
dw = np.concatenate((dw1.ravel(), db1, dw2.ravel(), db2), axis=0)
```

بهینه‌سازی پیش‌رفته

۵۵

□ دریافت پارامترها. پس از بهینه‌سازی، باید پارامترهای مختلف را از هم جدا کنیم.



$$F = 784$$

$$H = 20$$

$$C = 10$$

$$\begin{array}{ll} W^{(1)} \in \mathbb{R}^{F \times H} & b^{(1)} \in \mathbb{R}^H \\ W^{(2)} \in \mathbb{R}^{H \times C} & b^{(2)} \in \mathbb{R}^C \end{array}$$

$\leftarrow W$

```
w1 = np.reshape(W[: F * H], (F, H))
b1 = W[F * H: (F + 1) * H]
w2 = # ... get w2 and reshape it
b2 = # ... get b2
```

بهینه‌سازی پیش‌رفته: مراحل

۵۶

□ ماتریس‌ها و بردارهای $W^{(i)}$ و $b^{(i)}$ را ایجاد و به صورت تصادفی مقداردهی کنید.

```
# create and init parameters w1, w2  
w1 = np.random.randn(F, H) * 0.001  
w2 = np.random.randn(H, C) * 0.001  
  
b1 = np.zeros((H,))  
b2 = np.zeros((C,))
```

□ ماتریس‌ها و بردارهای $W^{(i)}$ و $b^{(i)}$ را به بردار W تبدیل کنید.

```
w = np.concatenate((w1.ravel(), b1, w2.ravel(), b2), axis=0)
```

بهینه‌سازی پیش‌رفته: مراحل

۵۷

- تابع بهینه‌سازی را به صورت زیر فراخوانی نمایید:

```
result = minimize(J, x0=W, args=(X_train, y_train), method='CG', jac=True)
```

- پس از بهینه‌سازی، مقدار پارامترها را به صورت زیر ذخیره کنید:

```
W = result.x
```

- ماتریس‌ها و بردارهای $W^{(i)}$ و $b^{(i)}$ را از بردار W استخراج کنید.

```
W1 = np.reshape(W[: F * H], (F, H))
```

```
b1 = W[F * H: (F + 1) * H]
```

```
W2 = # ... get W2 and reshape it
```

```
b2 = # ... get b2
```

بىرسى گرادىان

تَفْمِين عددي گراديانها

۵۹

□ مشتق تابع. در یک فضای ۱-بُعدی

$$\frac{df(x)}{dx} = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h}$$

□ در یک فضای چند بُعدی، گرادیان تابع، برداری است از مشتق‌های جزئی.

تَفْهِمِنِ عَدْدِي گُرَادِيَانِهَا

۶.

W

0.34
-1.11
0.78
0.12
0.55
2.81
-3.10
-1.50
0.33
...

Loss 1.25347

$W + h$

0.34	+ 0.0001
-1.11	
0.78	
0.12	
0.55	
2.81	
-3.10	
-1.50	
0.33	
...	

Loss 1.25322

dW

?
?
?
?
?
?
?
?
?
?
...

تَفْهِين عَدْدِي گَرَادِيانُهَا

۶۱

W

0.34
-1.11
0.78
0.12
0.55
2.81
-3.10
-1.50
0.33
...

Loss 1.25347

$W + h$

0.34	+ 0.0001
-1.11	
0.78	
0.12	
0.55	
2.81	
-3.10	
-1.50	
0.33	
...	

Loss 1.25322

dW

-2.50
?
?
?

$$(1.25322 - 1.25347) / 0.0001 = -2.5$$

$$\frac{df(x)}{dx} = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h}$$

?
...

تَفْهِمِنِ عَدْدِي گُرَادِيَانِهَا

۶۲

W

0.34
-1.11
0.78
0.12
0.55
2.81
-3.10
-1.50
0.33
...

Loss 1.25347

$W + h$

0.34
-1.11
0.78
0.12
0.55
2.81
-3.10
-1.50
0.33
...

Loss 1.25353

dW

-2.50
?
?
?
?
?
?
?
?
...

تَفْهِين عَدْدِي گِرَادِيان‌ها

٦٣

W

0.34
-1.11
0.78
0.12
0.55
2.81
-3.10
-1.50
0.33
...

Loss 1.25347

$W + h$

0.34
-1.11
0.78
0.12
0.55
2.81
-3.10
-1.50
0.33
...

Loss 1.25353

dW

-2.50
0.60
?
?

$$(1.25353 - 1.25347) / 0.0001 = 0.6$$

$$\frac{df(x)}{dx} = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

?
...

تَفْهِمِنِ عَدْدِي گُرَادِيَانِهَا

۶۴

W

0.34
-1.11
0.78
0.12
0.55
2.81
-3.10
-1.50
0.33
...

Loss 1.25347

$W + h$

0.34
-1.11
0.78
0.12
0.55
2.81
-3.10
-1.50
0.33
...

+ 0.0001

Loss 1.25347

dW

-2.50
0.60
?
?
?
?
?
?
?
...

تَفْهِين عَدْدِي گَرَادِيانُهَا

٦٨

W

0.34
-1.11
0.78
0.12
0.55
2.81
-3.10
-1.50
0.33
...

Loss 1.25347

$W + h$

0.34
-1.11
0.78
0.12
0.55
2.81
-3.10
-1.50
0.33
...

Loss 1.25347

dW

-2.50
0.60
0.00
?
?

$$(1.25347 - 1.25347) / 0.0001 = 0.0$$

$$\frac{df(x)}{dx} = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

تَفْمِين عَدْدِي گَرَادِيَانُهَا

۶۶

```
def eval_numerical_gradient(f, x):

    fx = f(x)
    grad = np.zeros(x.shape)
    h = 0.00001

    it = np.nditer(x, flags=[‘multi_index’], op_flags=[‘readwrite’])
    while not it.finished:

        ix = it.multi_index
        old_value = x[ix]
        x[ix] += h
        fxh = f(x) # evaluate f(x + h)
        x[ix] = old_value

        grad[ix] = (fxh - fx) / h # compute the partial derivative
        it.iternext() # step to next dimension

    return grad
```

$$\frac{df(x)}{dx} = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h}$$

معایب.

تقریبی

بسیار زمانبر

بررسی گرادیان

۶۷

□ به طور خلاصه.

□ گرادیان عددی: تقریبی، زمانبر، پیاده‌سازی آسان!

□ گرادیان تحلیلی: دقیق، سریع، امکان بروز خطا در پیاده‌سازی!

□ در عمل.

□ همیشه از گرادیان تحلیلی استفاده می‌کنیم.

□ اما به منظور اطمینان از درستی پیاده‌سازی، گرادیان تحلیلی را با گرادیان عددی مقایسه می‌کنیم.

بررسی گرادیان



کرادیان کاہشی با دسنهای کوپک

الگوریتم گرادیان کاهشی

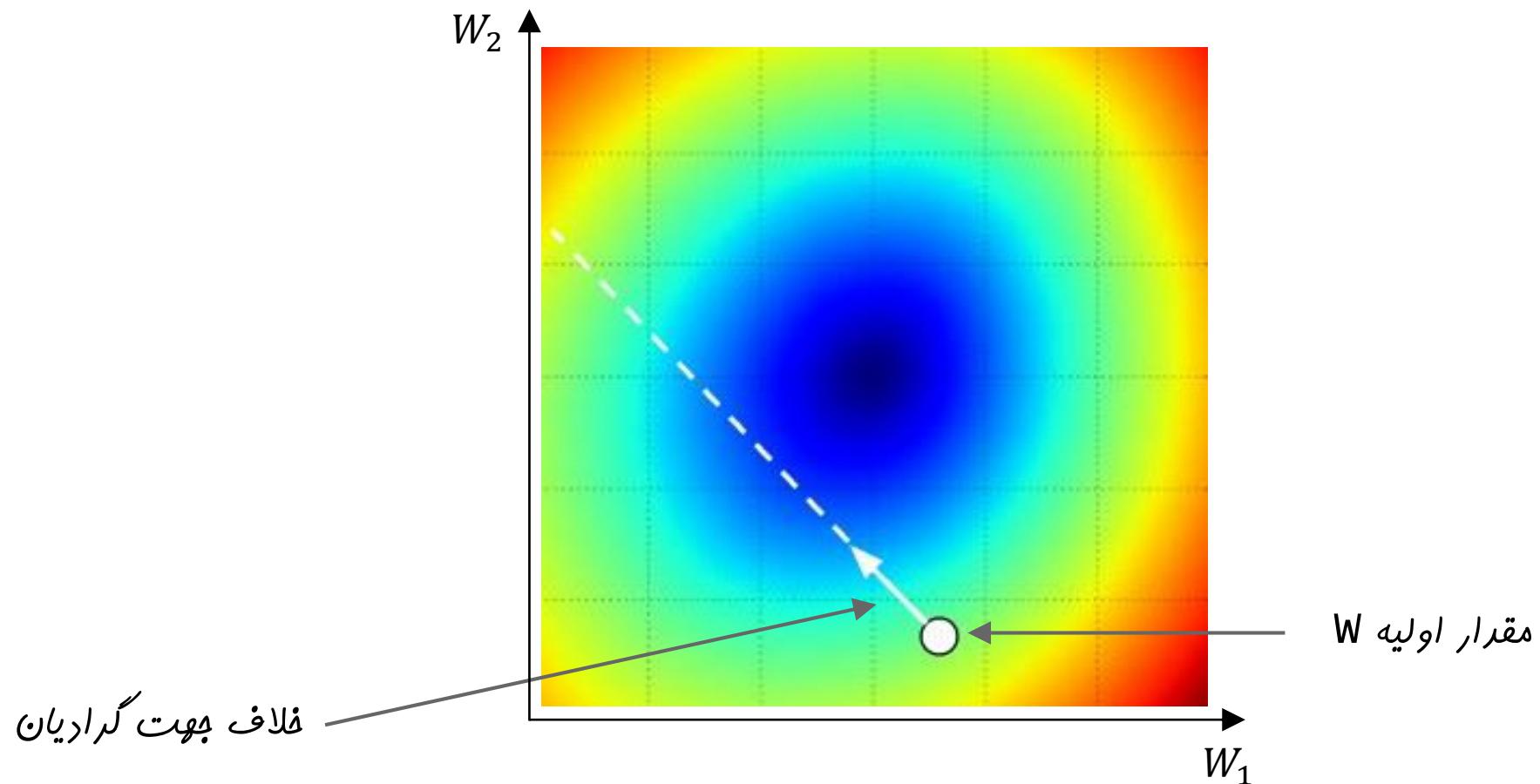
۶۹

```
# Vanilla Gradient Descent

while True:
    gradient = evaluate_gradient(loss_fun, data, weights)
    weights += -step_size * gradient # weight update
```

الگوریتم گرادیان کاهشی

۷۰



یک نسخه کاراَر از الگوریتم گرادیان کاهشی

۷۱

- گرادیان کاهشی با دسته‌های کوچک.
- برای محاسبه گرادیان تابع هزینه در هر تکرار، تنها از بخش کوچکی از داده‌های آموزشی استفاده کن.

```
# Mini-batch Gradient Descent
```

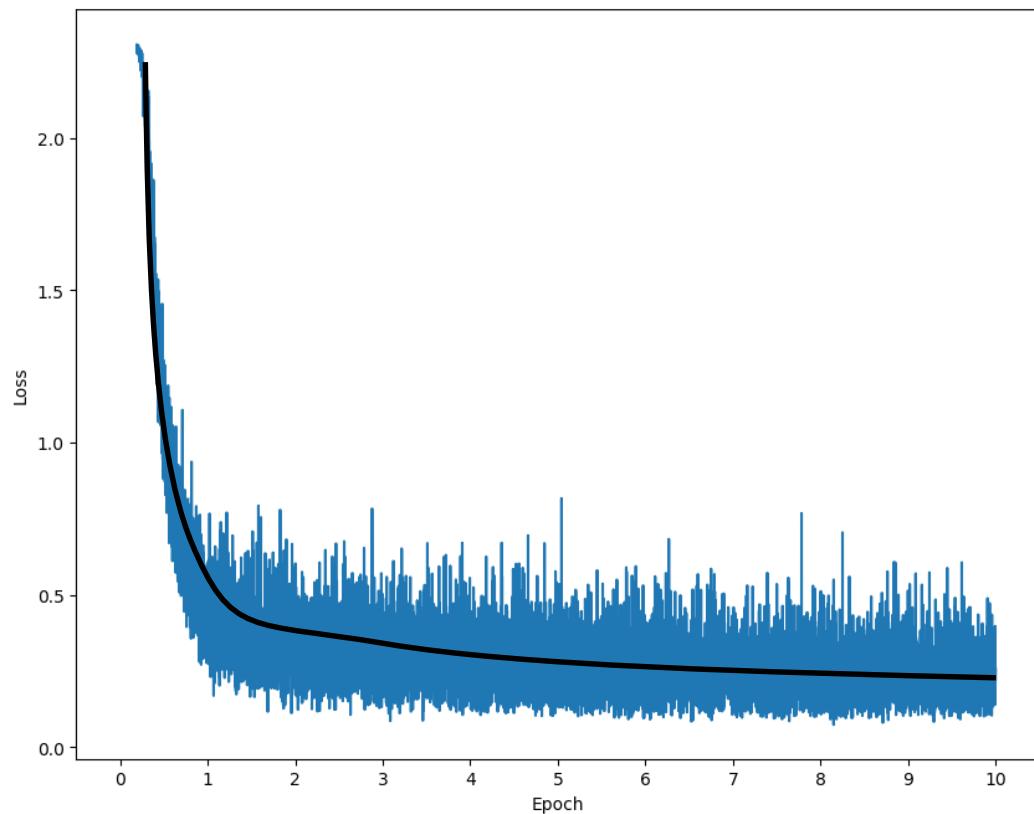
```
while True:  
    data_batch = sample_training_data(data, 256) # sample 256 examples  
    gradient = evaluate_gradient(loss_fun, data_batch, weights)  
    weights += -step_size * gradient # weight update
```

- مقادیر متداول برای اندازه دسته: ۳۲، ۶۴، ۱۲۸ و ۲۵۶.

گرادیان کاهشی با دسته‌های کوچک

۷۲

□ نمونه اجرای الگوریتم گرادیان کاهشی با دسته‌های کوچک در یک شبکه عصبی.



هزینه در طول زمان کاهش می‌یابد.

ماشین‌های بردار پشتیبان

سید ناصر رضوی www.snrazavi.ir

۱۳۹۷

فهرست مطالب

۲

- انگیزه. مرز تصمیم‌گیری بهینه
- مفاهیم پایه. بردارهای پشتیبان و بیشینه‌سازی حاشیه
- تابع هدف. مسئله اصلی و مسئله دوگان
- دسته‌بندی خطی و غیرخطی. حاشیه نرم
- دسته‌بندی غیرخطی. ترفندهای کرنل
- دسته‌بندی چند دسته‌ای. ماشین بردار پشتیبان چند دسته‌ای

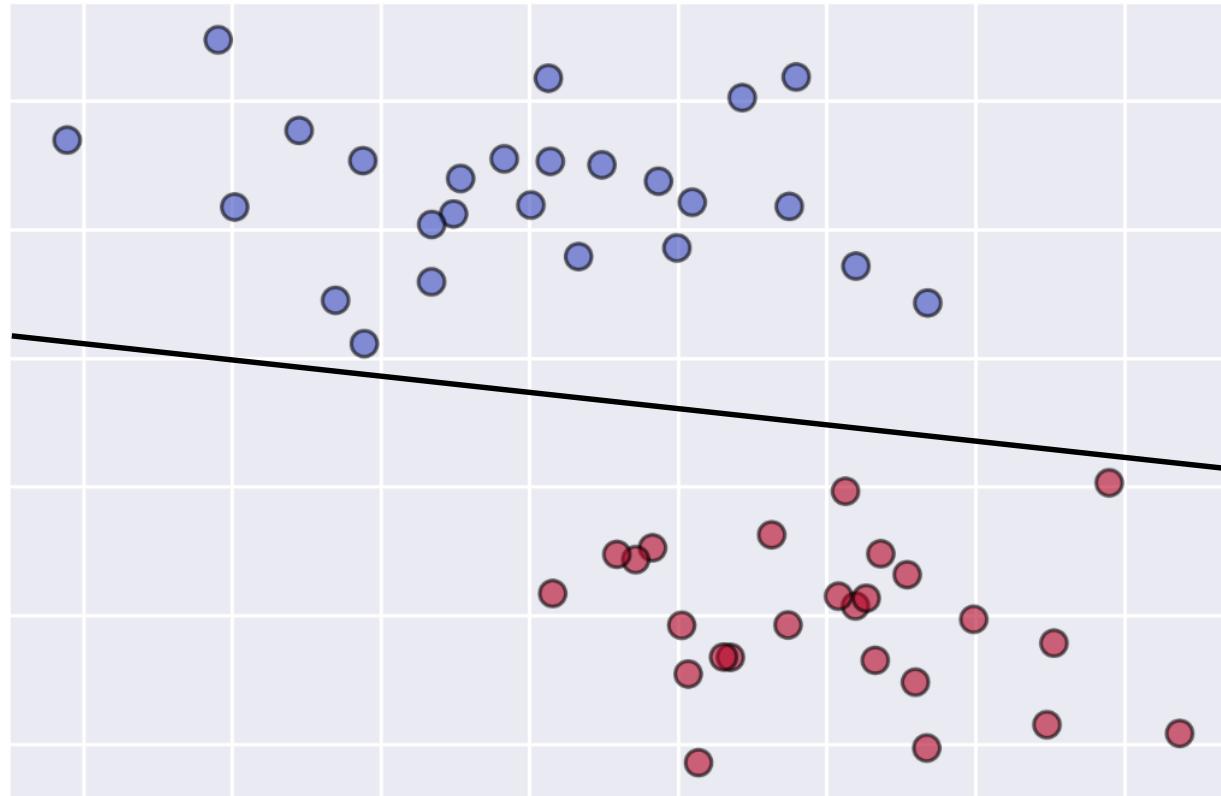
معرفی

۳

- ماشین‌های بردار پشتیبان [وپنیک، ۱۹۹۲]
- یکی از پرطرفدارترین الگوریتم‌های یادگیری ماشین!
- جداسازی بهتر داده‌ها نسبت به سایر روش‌های یادگیری ماشین (مسائل دسته‌بندی)!
- استفاده از آن نسبتاً آسان است!
- استفاده از ترفند کرنل:
- دسته‌بندی، رگرسیون، تخمین توزیع، دسته‌بندی تک دسته‌ای و ...

انگیزه: داده‌های تفکیک‌پذیر فطی

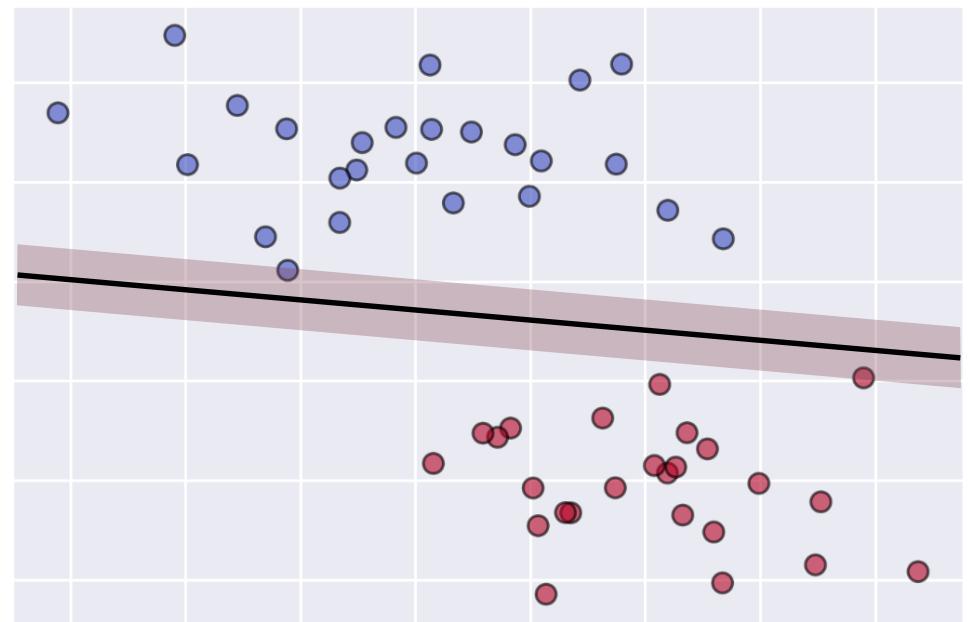
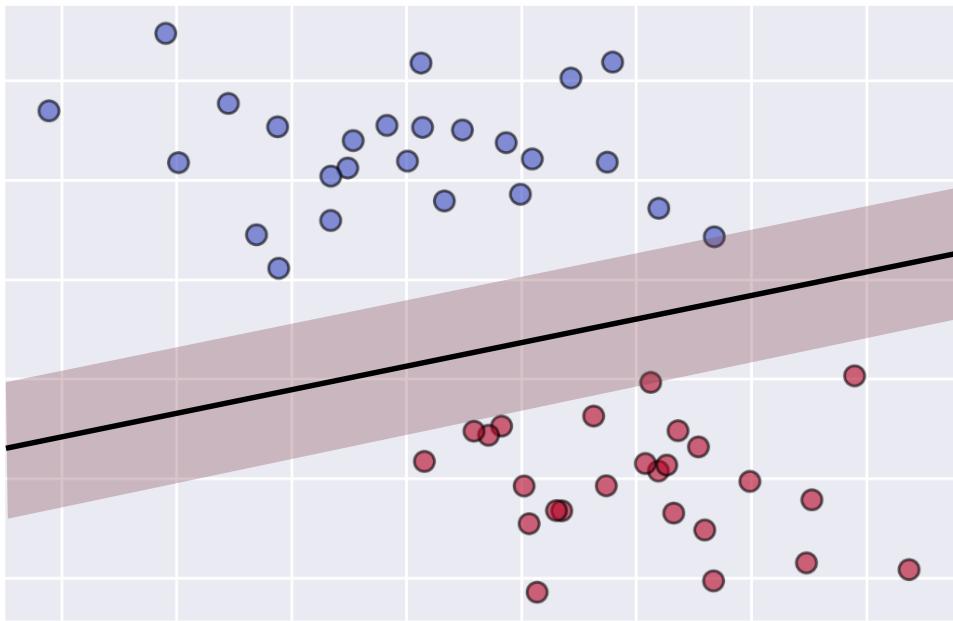
۴



انگیزه: مرز تصمیم‌گیری بهینه

۵

□ پرسش. کدام مرز تصمیم‌گیری بهتر است؟

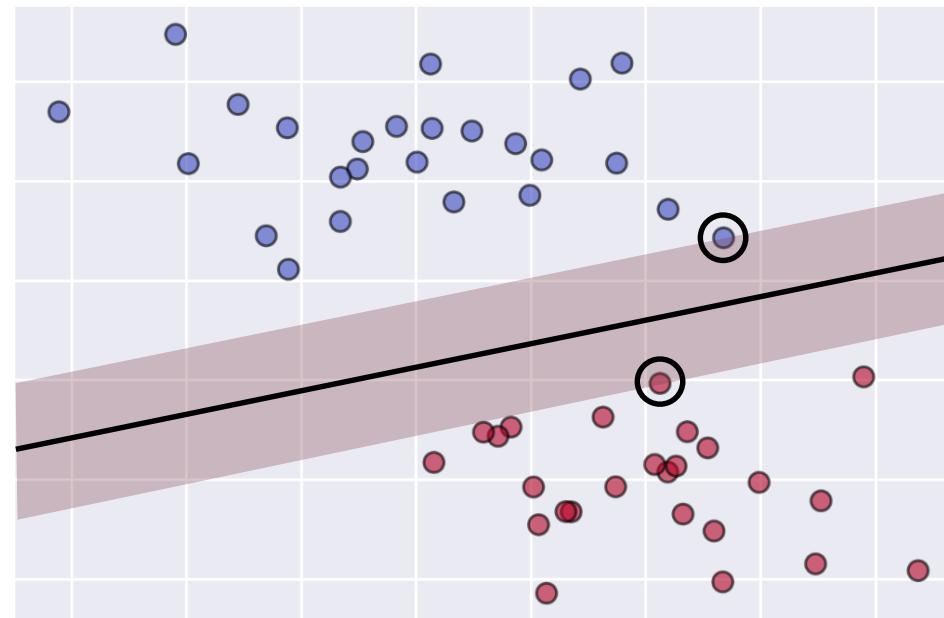


□ راه حل بیشترین حاشیه. بیشترین پایداری در برابر تخریب داده‌ها. [افزایش قابلیت تعمیم]

انگیزه: بردارهای پشتیبان

۶

□ بردار پشتیبان. نزدیکترین داده‌ها به مرز تصمیم‌گیری.

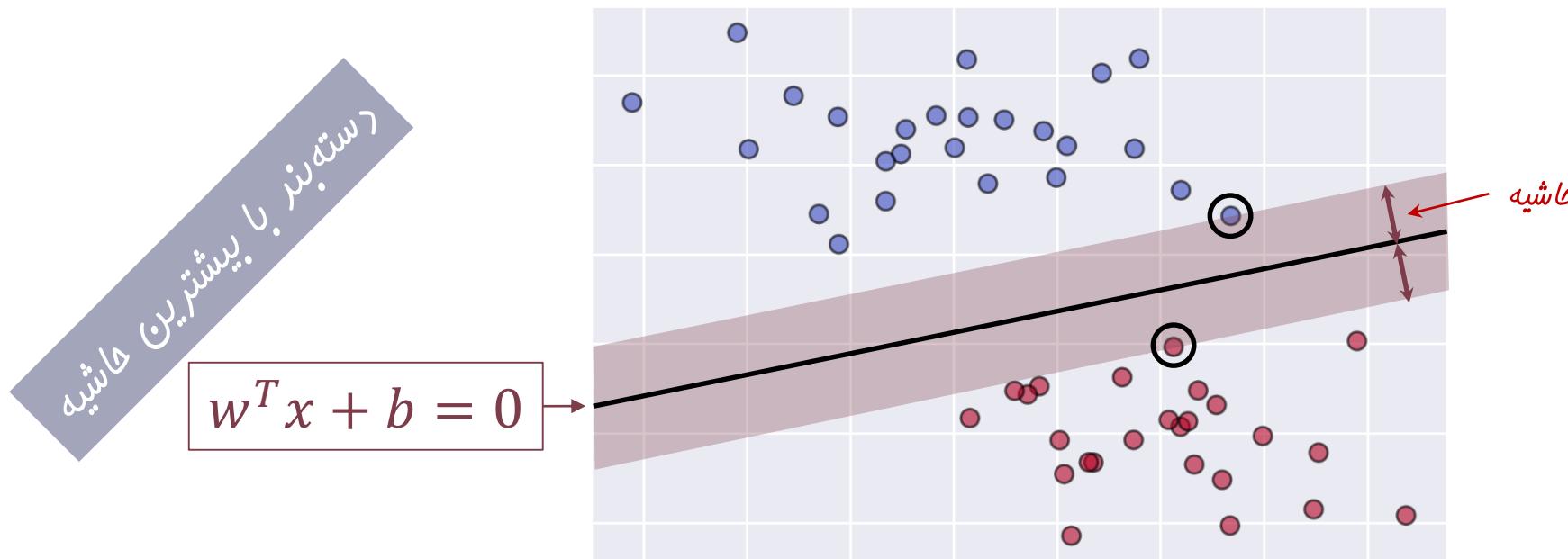


□ هدف. بیشینه کردن فاصله بردارهای پشتیبان از مرز تصمیم‌گیری.

انگیزه: بیشینه‌سازی حاشیه

۷

□ **حاشیه.** فاصله بردارهای پشتیبان تا مرز تصمیم‌گیری.



□ **هدف.** بیشینه کردن فاصله بردارهای پشتیبان از مرز تصمیم‌گیری.

درز تصمیمگیری بهینه: نمادها

۸

نمونه‌های آموزشی. □

$$X = (\mathbf{x}^t, y^t),$$

$$y^t = \begin{cases} +1 & \text{if } x^t \in C_1 \\ -1 & \text{if } x^t \in C_2 \end{cases}$$

هدف. یافتن بردار w و مقدار b به گونه‌ای که:

$$w^T x^t + b \geq +1 \quad \text{for } y^t = +1$$

$$w^T x^t + b \leq -1 \quad \text{for } y^t = -1$$

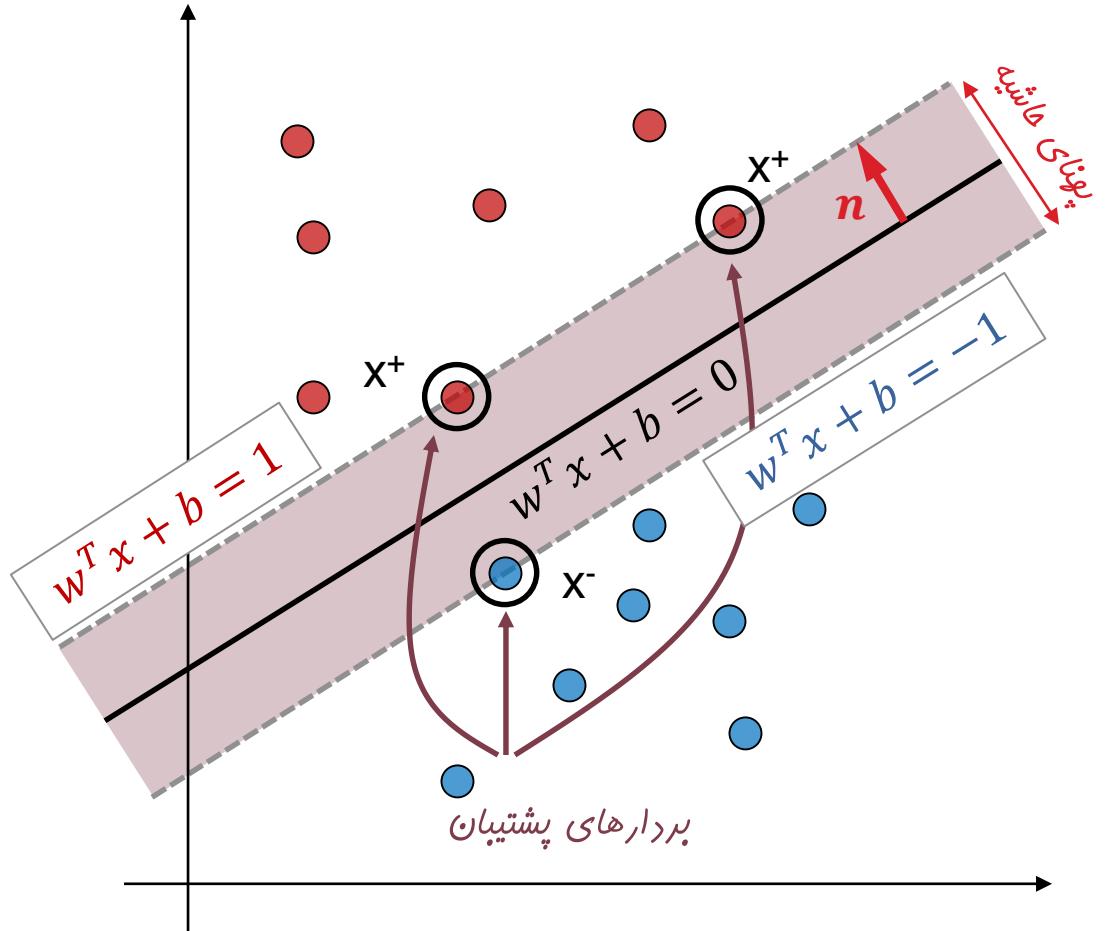


$$y^t(w^T x^t + b) \geq +1$$

$$\max(0, 1 - y^t(w^T x^t + b))$$

تابع هدف: محاسبه حاشیه

۹



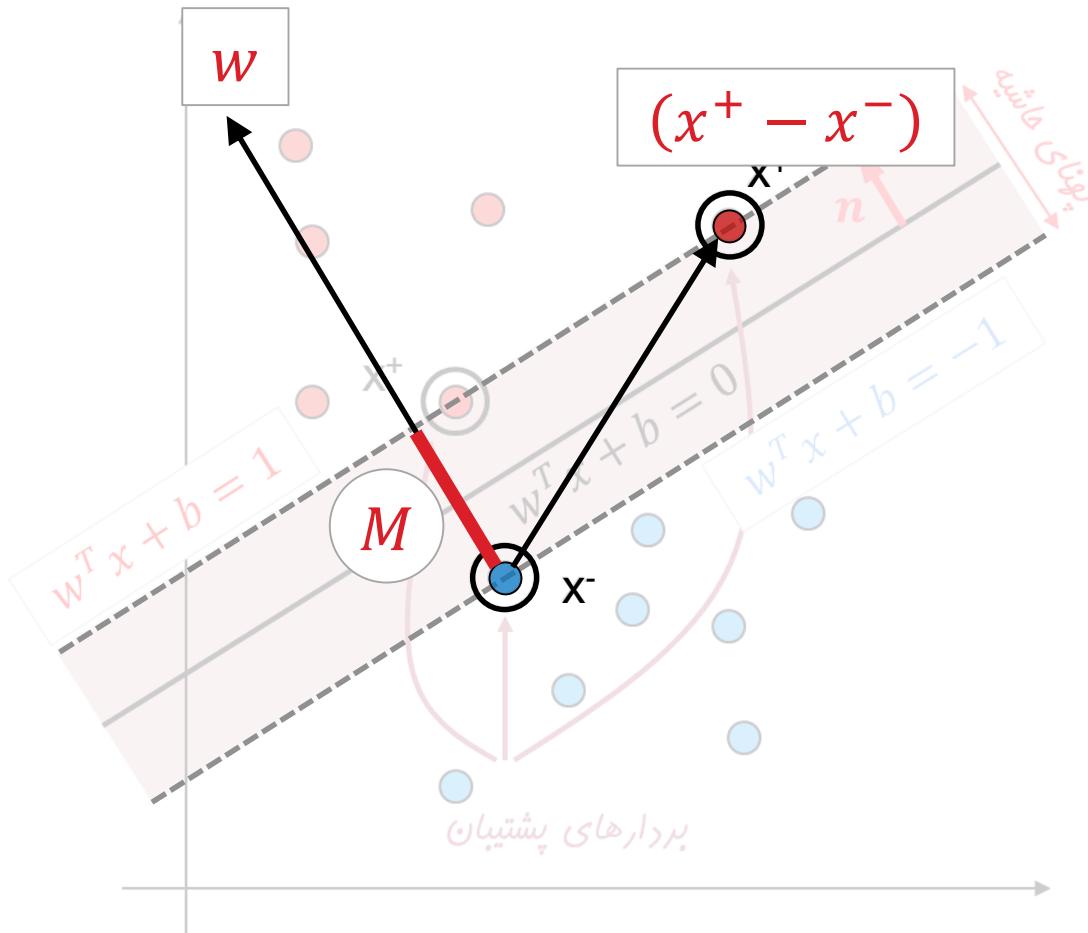
می‌دانیم:

$$w^T x^+ + b = +1$$

$$w^T x^- + b = -1$$

تابع هدف: محاسبه حاشیه

۱۰



می‌دانیم:

$$w^T x^+ + b = +1$$

$$w^T x^- + b = -1$$

بنابراین:

$$w^T(x^+ - x^-) = 2$$

$$\Rightarrow \|w\| \cdot \|x^+ - x^-\| \cos \alpha = 2$$

$$\Rightarrow \|w\| \cdot M = 2 \Rightarrow \boxed{M = 2/\|w\|}$$

تابع هدف

۱۱

□ هدف. بیشینه کردن اندازه حاشیه [فاصله بردارهای پشتیبان از مرز تصمیم‌گیری].

$$M = \frac{2}{\|w\|}$$

□ توجه. برای بیشینه کردن حاشیه، می‌توان اندازه بردار w را کمینه نمود.

□ محدودیت‌ها. مرز تصمیم‌گیری باید داده‌های هر دو دسته را به درستی از یکدیگر تفکیک کند.

تابع هدف: بیان (سمی)

۱۲

تابع هدف. □

$$\min \frac{1}{2} \|\mathbf{w}\|^2$$

s.t. $(\mathbf{w}^T \mathbf{x}^t + b) \geq +1 \quad if \ y^t = +1$

$$(\mathbf{w}^T \mathbf{x}^t + b) \leq -1 \quad if \ y^t = -1$$

ساده‌سازی. □

$$\min \frac{1}{2} \|\mathbf{w}\|^2$$

s.t. $y^t (\mathbf{w}^T \mathbf{x}^t + b) \geq +1$

تابع هدف: بیان (سمی)

۱۳

$$\begin{aligned} & \min \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t. } & y^t(\mathbf{w}^T \mathbf{x}^t + b) \geq +1 \end{aligned}$$

بینه‌سازی مدرس

□ تابع هدف.

□ حل مسئله با استفاده از ضرایب لاغرانژ.

$$\begin{aligned} L_p &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{t=1}^m \alpha^t [y^t(\mathbf{w}^T \mathbf{x}^t + b) - 1] \\ &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{t=1}^m \alpha^t y^t (\mathbf{w}^T \mathbf{x}^t + b) + \sum_{t=1}^m \alpha^t \end{aligned}$$



ژوزف لویی لاغرانژ

تابع هدف: بیان (سمی)

۱۴

تابع هدف. □

$$\begin{aligned}L_p &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{t=1}^m \alpha^t [y^t (\mathbf{w}^T \mathbf{x}^t + b) - 1] \\&= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{t=1}^m \alpha^t y^t (\mathbf{w}^T \mathbf{x}^t + b) + \sum_{t=1}^m \alpha^t\end{aligned}$$

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{t=1}^m \alpha^t y^t \mathbf{x}^t$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{t=1}^m \alpha^t y^t = 0$$

مرز تضمین‌گیری یک ترکیب فطی از داره‌های آموزشی

تابع هدف: شکل دوگان

۱۵

$$\begin{aligned}
 L_d &= \frac{1}{2}(\mathbf{w}^T \mathbf{w}) - \mathbf{w}^T \sum_{t=1}^m \alpha^t y^t \mathbf{x}^t - b \sum_{t=1}^m \alpha^t y^t + \sum_{t=1}^m \alpha^t \\
 &= -\frac{1}{2}(\mathbf{w}^T \mathbf{w}) + \sum_{t=1}^m \alpha^t \\
 &= -\frac{1}{2} \sum_{t=1}^m \sum_{s=1}^m \alpha^t \alpha^s y^t y^s (\mathbf{x}^t)^T \mathbf{x}^s + \sum_{t=1}^m \alpha^t
 \end{aligned}$$

الگوریتم
بینه‌سازی ترتیبی مینیمال
پلت (۱۹۹۹)

تابع هدف. □

subject to $\sum_{t=1}^m \alpha^t y^t = 0$ and $\alpha^t \geq 0 \forall t$

- مقدار بسیاری از ضرایب آلفا برابر با صفر است و تنها تعداد اندکی دارای مقدار بزرگ‌تر از صفر هستند؛
- داده‌هایی که به ازای آنها مقدار آلفا بزرگ‌تر از صفر است، همان **بردارهای پشتیبان** هستند.

تابع هدف: شکل برداری

۱۶

□ تابع هدف.

$$\begin{aligned} L_d &= -\frac{1}{2} \sum_{t=1}^m \sum_{s=1}^m \alpha^t \alpha^s y^t y^s (\mathbf{x}^t)^T \mathbf{x}^s + \sum_{t=1}^m \alpha^t \\ &= -\frac{1}{2} \alpha^T Q \alpha + e^T \alpha \end{aligned}$$

الگوریتم
بهینه‌سازی ترتیبی مینیمال
پلت (۱۹۹۹)

$$Q_{ts} = y^t y^s (\mathbf{x}^t)^T \mathbf{x}^s, \quad e = [1 \ 1 \ \dots \ 1]^T \in \mathbb{R}^m$$

subject to $\sum_{t=1}^m \alpha^t y^t = 0$ and $\alpha^t \geq 0 \ \forall t$

- مقدار بسیاری از ضرایب آلفا برابر با صفر است و تنها تعداد اندکی دارای مقدار بزرگ‌تر از صفر هستند؛
- داده‌هایی که به ازای آنها مقدار آلفا بزرگ‌تر از صفر است، همان **بردارهای پشتیبان** هستند.

داده‌های تفکیک‌ناپذیر خطی: حاشیه نرم

۱۷

$$y^t(\mathbf{w}^T \mathbf{x}^t + b) \geq 1 - \varepsilon^t$$

$$\text{soft error} = \sum_{t=1}^m \varepsilon^t$$

ضریب بریمه

$$\begin{aligned} & \min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{t=1}^m \varepsilon^t \\ \text{s.t. } & y^t(\mathbf{w}^T \mathbf{x}^t + b) \geq 1 - \varepsilon^t \\ & \varepsilon^t \geq 0 \end{aligned}$$

□ حاشیه نرم. اجازه دادن اندکی خطا در جداسازی!

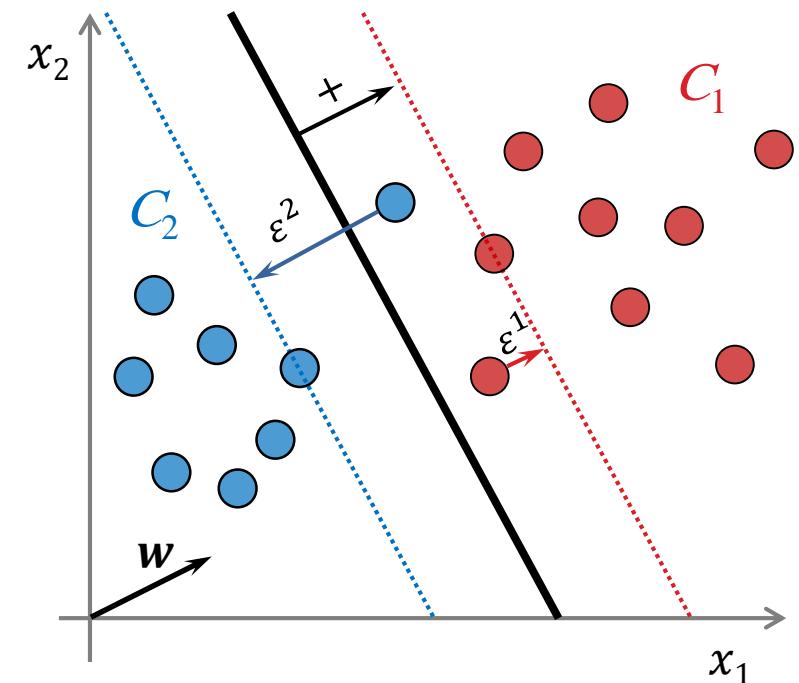
□ خطای نرم.

□ تابع هدف جدید.

داده‌های تفکیک‌ناپذیر خطي: حاشیه نرخ

۱۸

$$\begin{aligned} & \min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{t=1}^m \varepsilon^t \\ \text{s.t. } & y^t (\mathbf{w}^T \mathbf{x}^t + b) \geq 1 - \varepsilon^t \\ & \varepsilon^t \geq 0 \end{aligned}$$



ضد ایب لاغرانژ

$$L_p = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{t=1}^m \varepsilon^t - \sum_{t=1}^m \alpha^t [y^t (\mathbf{w}^T \mathbf{x}^t + b) - 1 + \varepsilon^t] - \sum_{t=1}^m \mu^t \varepsilon^t$$

داده‌های تفکیک‌ناپذیر خطي: ماشین نرخ

۱۹

$$L_p = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{t=1}^m \varepsilon^t - \sum_{t=1}^m \alpha^t [y^t (\mathbf{w}^T \mathbf{x}^t + b) - 1 + \varepsilon^t] - \sum_{t=1}^m \mu^t \varepsilon^t$$

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{t=1}^m \alpha^t y^t \mathbf{x}^t$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{t=1}^m \alpha^t y^t = 0$$

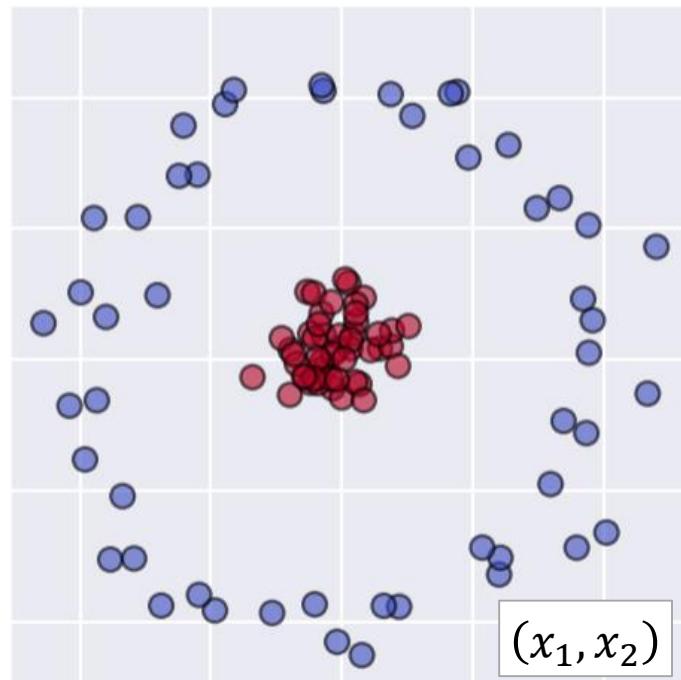
$$\frac{\partial L_p}{\partial \varepsilon^t} = 0 \Rightarrow C - \alpha^t - \mu^t = 0 \Rightarrow 0 \leq \alpha^t \leq C$$

ترفند کرنل و دسَبندی غیرخطی

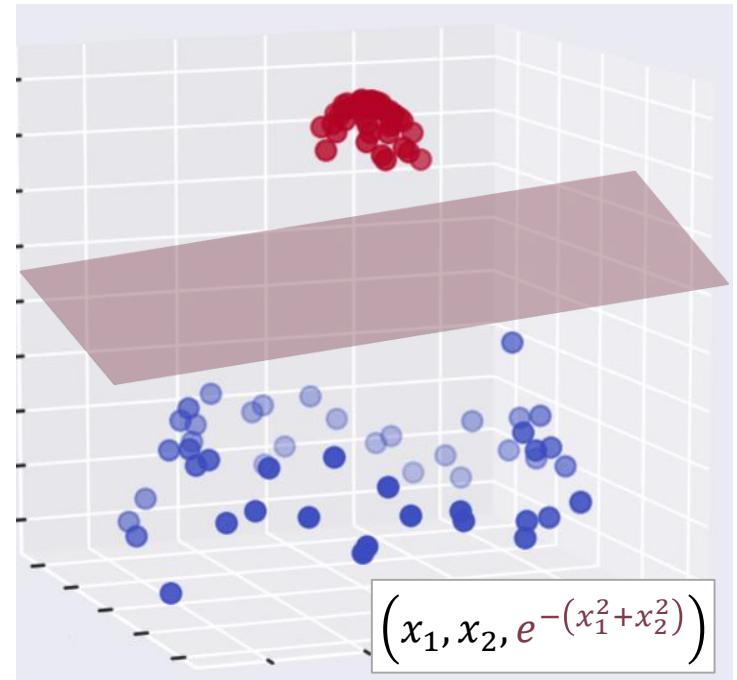
توابع کرnel

۲۱

- ایده. نگاشت مسئله به یک فضای ویژگی جدید با استفاده از تبدیلات غیرخطی.
- استفاده از یک مدل خطی در فضای جدید به منظور دسته‌بندی داده‌ها.



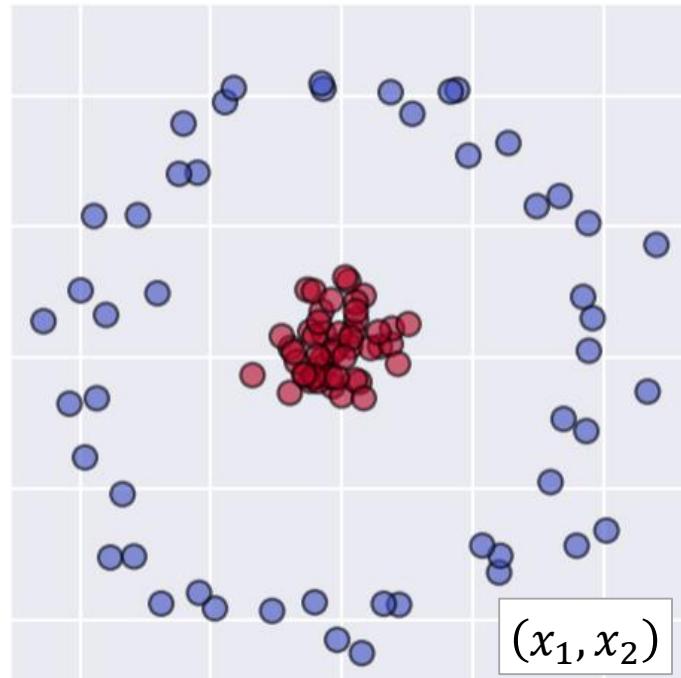
$$\Phi: x \rightarrow \varphi(x)$$



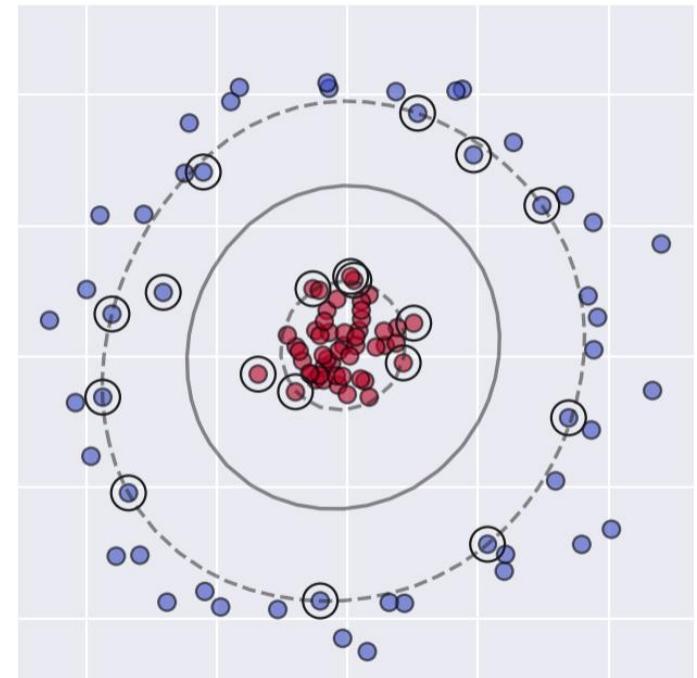
توابع کرnel

۲۲

- ایده. نگاشت مسئله به یک فضای ویژگی جدید با استفاده از تبدیلات غیرخطی.
- استفاده از یک مدل خطی در فضای جدید به منظور دسته‌بندی داده‌ها.
- مدل خطی در فضای جدید متناظر با یک مدل غیرخطی در فضای اصلی است.

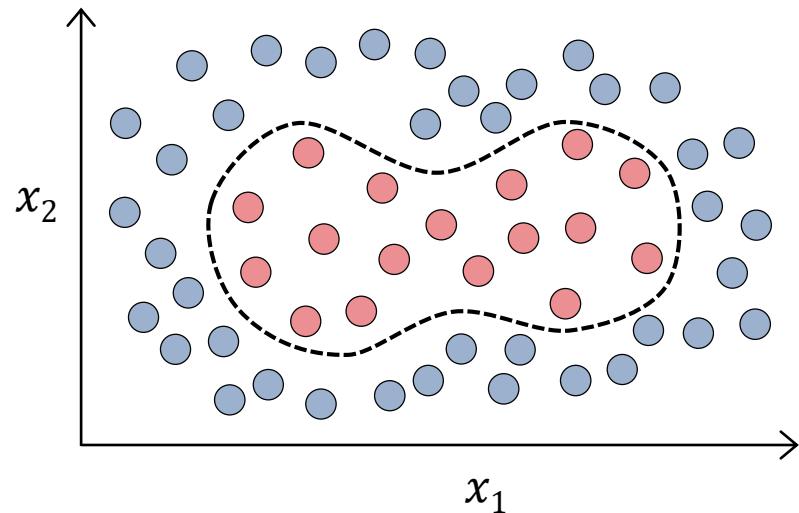


$$\Phi: x \rightarrow \varphi(x)$$



مرز تضمین‌گیری غیرخطی

۲۳



پیش‌بینی. اگر $y = 1$: □

$$h(x) = b + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2 + w_5x_1x_2 + \dots \geq 0$$

ویژگی‌ها. □

$$f_1 = x_1, \quad f_2 = x_2, \quad f_3 = x_1^2, \quad f_4 = x_2^2, \quad f_5 = x_1x_2, \quad \dots$$

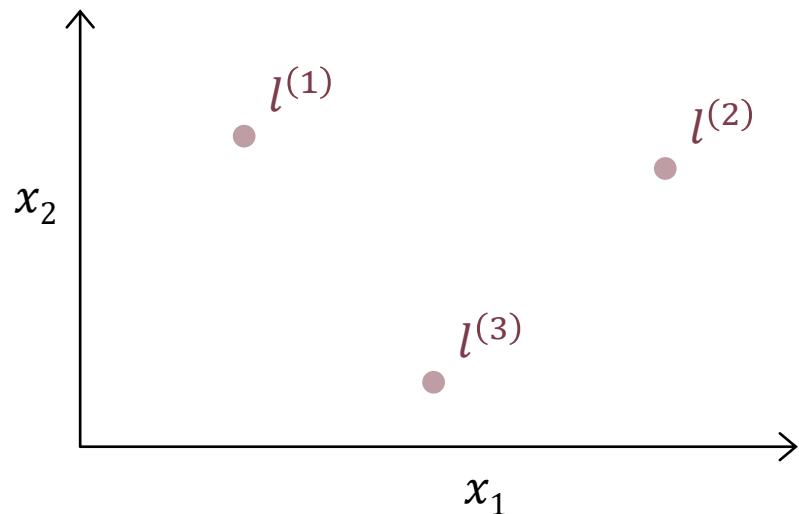
$$h(f) = b + w_1f_1 + w_2f_2 + w_3f_3 + w_4f_4 + w_5f_5 + \dots \quad \leftarrow \text{مرز تضمین‌گیری خطی}$$

پرسش. آیا روش بهتری برای انتخاب ویژگی‌ها وجود دارد? □

کرنل‌ها به عنوان معیار شباهت

۲۴

□ ایده. با داشتن x , مجموعه جدید ویژگی‌ها را بر اساس **شباهت** آن با نقاط راهنمای $l^{(1)}$, $l^{(2)}$ و $l^{(3)}$ انتخاب کن.



$$f_1 = sim(x, l^{(1)}) = \exp\left(-\frac{\|x - l^{(1)}\|^2}{2\sigma^2}\right)$$

$$f_2 = sim(x, l^{(2)}) = \exp\left(-\frac{\|x - l^{(2)}\|^2}{2\sigma^2}\right)$$

$$f_3 = sim(x, l^{(3)}) = \exp\left(-\frac{\|x - l^{(3)}\|^2}{2\sigma^2}\right)$$

کرنل (کرنل گوسی)

□ تابع کرنل. معیاری به منظور محاسبه شباهت میان داده‌های x و y

کرنل‌ها به عنوان معیار شباهت

۲۵

□ تابع کرنل.

$$f_i = sim(x, l^{(i)}) = \exp\left(-\frac{\|x - l^{(i)}\|^2}{2\sigma^2}\right)$$

□ حالت اول. $x \approx l^{(i)}$.

$$f_i \approx \exp\left(-\frac{0}{2\sigma^2}\right) = \exp(0) = 1$$

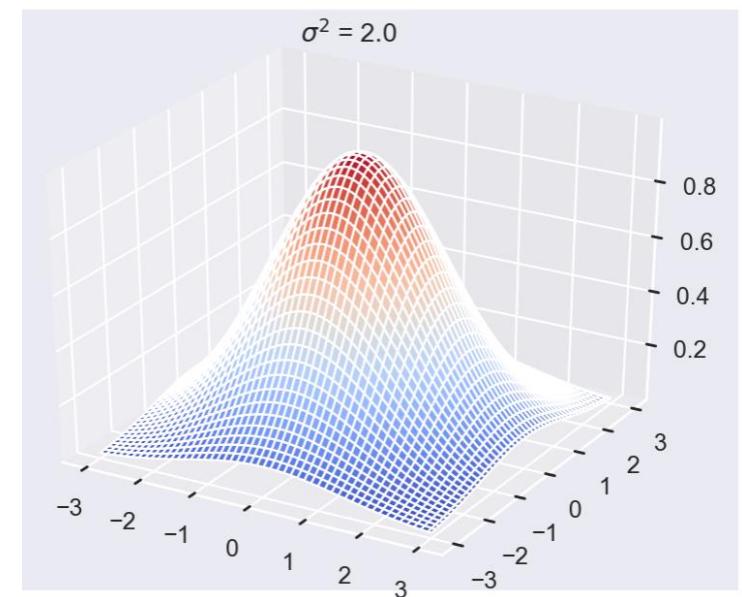
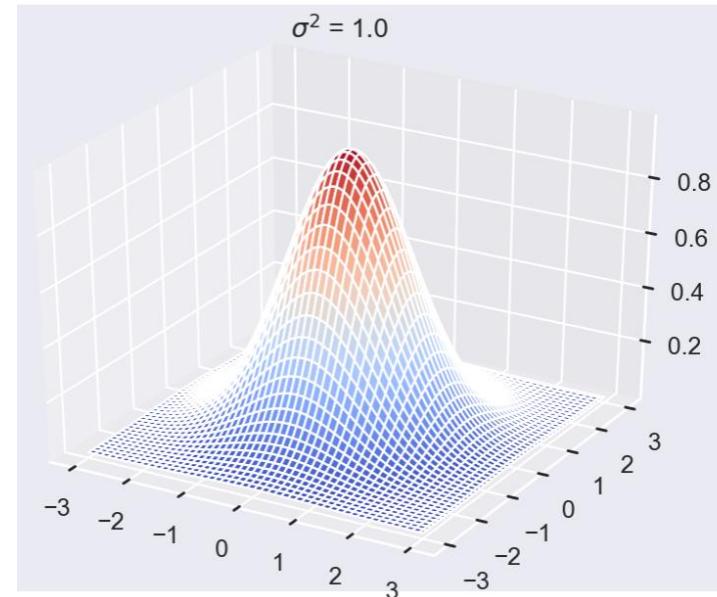
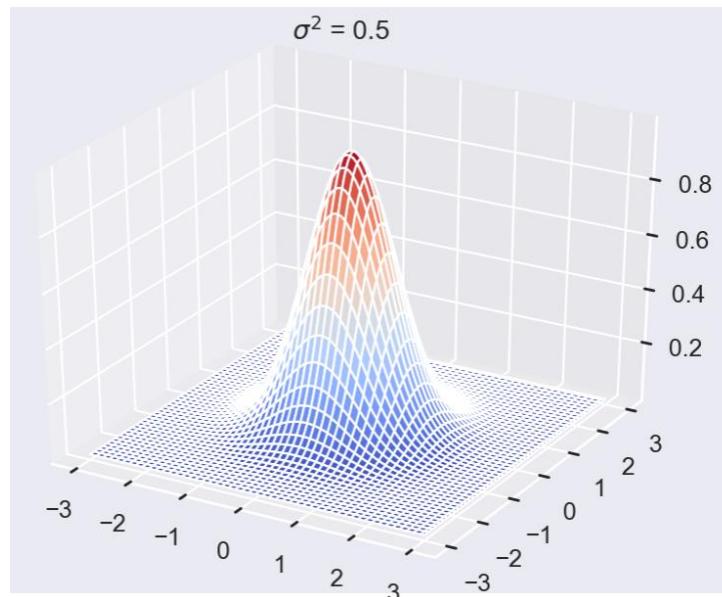
□ حالت دوم. x بسیار دور از $l^{(i)}$

$$f_i \approx \exp\left(-\frac{\infty}{2\sigma^2}\right) = \exp(-\infty) = 0$$

کرنل‌ها به عنوان معیار شباهت

۲۶

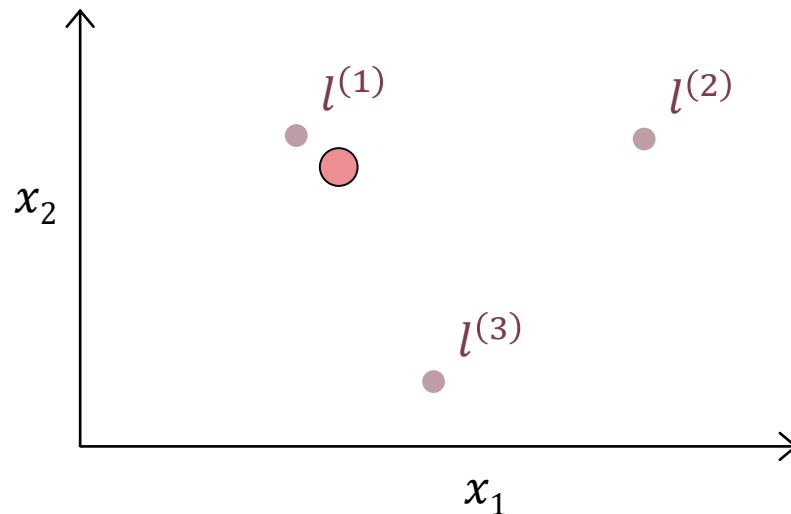
$$f_i = \text{sim}(x, l^{(i)}) = \exp\left(-\frac{\|x - l^{(i)}\|^2}{2\sigma^2}\right)$$



کرنل‌ها به عنوان معیار شباهت

۲۷

پیش‌بینی. $y = 1$ اگر: □



$$b + w_1 f_1 + w_2 f_2 + w_3 f_3 \geq 0$$

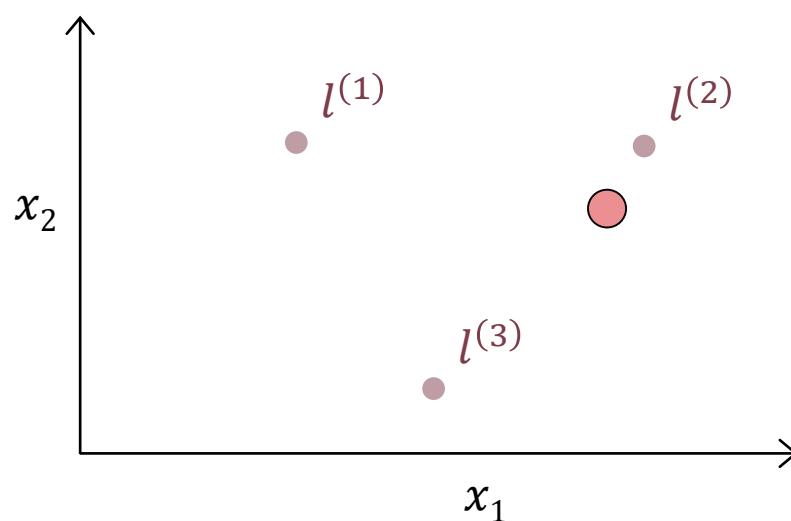
$$\begin{array}{cccc} \uparrow & \uparrow & \uparrow & \uparrow \\ -0.5 & 1.0 & 1.0 & 0.0 \end{array}$$

$$f_1 \approx 1, f_2 \approx f_3 \approx 0$$

$$h(f) \approx -0.5 + (1.0)(1.0) + (1.0)(0.0) + (0.0)(0.0) = 0.5 \geq 0 \Rightarrow \boxed{y = 1}$$

کرنل‌ها به عنوان معیار شباهت

۲۸



اگر $y = 1$ پیش‌بینی. \square

$$b + w_1 f_1 + w_2 f_2 + w_3 f_3 \geq 0$$

$$\begin{array}{cccc} \uparrow & \uparrow & \uparrow & \uparrow \\ -0.5 & 1.0 & 1.0 & 0.0 \end{array}$$

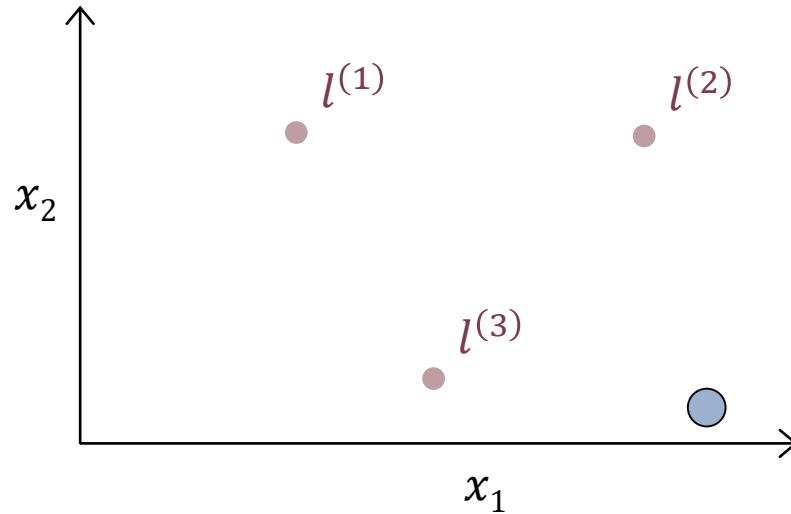
$$f_1 \approx f_3 \approx 0, f_2 \approx 1$$

$$h(f) \approx -0.5 + (1.0)(0.0) + (1.0)(1.0) + (0.0)(0.0) = 0.5 \geq 0 \Rightarrow y = 1$$

کرنل‌ها به عنوان معیار شباهت

۲۹

پیش‌بینی. $y = 1$ اگر: □



$$b + w_1 f_1 + w_2 f_2 + w_3 f_3 \geq 0$$

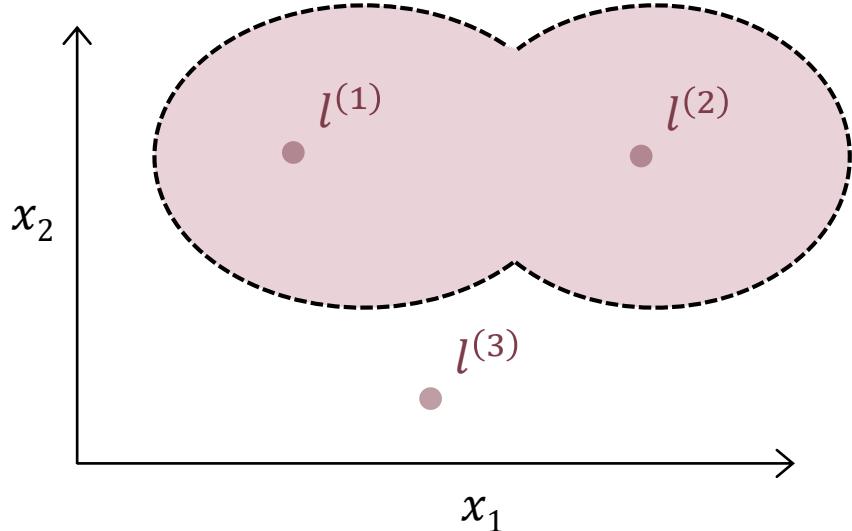
$$\begin{array}{cccc} \uparrow & \uparrow & \uparrow & \uparrow \\ -0.5 & 1.0 & 1.0 & 0.0 \end{array}$$

$$f_1 \approx f_2 \approx f_3 \approx 0$$

$$h(f) \approx -0.5 + (1.0)(0.0) + (1.0)(0.0) + (0.0)(0.0) = -0.5 < 0 \Rightarrow y = 0$$

کرنل‌ها به عنوان معیار شباهت

۳۰



پیش‌بینی. $y = 1$ اگر: □

$$b + w_1f_1 + w_2f_2 + w_3f_3 \geq 0$$

\uparrow \uparrow \uparrow \uparrow
-0.5 1.0 1.0 0.0

مرز تصمیم‌گیری. نقاط نزدیک به $l^{(1)}$ و $l^{(2)}$ را در دسته ۱ و سایر نقاط را در دسته صفر دسته‌بندی می‌کند. □

چند پرسش

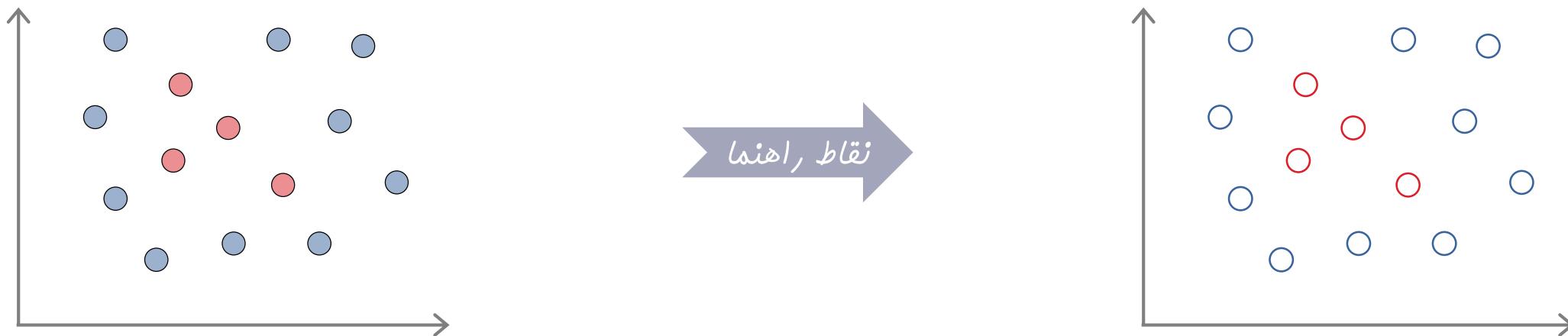
۳۱

- الگوریتم یادگیری نقاط راهنمای چگونه به صورت خودکار انتخاب می‌کند؟
- مقدار مناسب برای پارامترهای تابع کرنل چگونه تعیین می‌شوند؟
- آیا انواع دیگری از کرنل‌ها وجود دارد؟

انتخاب نقاط راهنمایی

۳۲

- الگوریتم یادگیری نقاط راهنمایی را چگونه به صورت خودکار انتخاب می‌کند؟
- به ازای هر نمونه در مجموعه آموزشی، یک نقطه راهنمایی مساوی با آن نمونه انتخاب می‌شود.



نگاشت ویژگی‌ها

۳۳

□ مجموعه آموزشی.

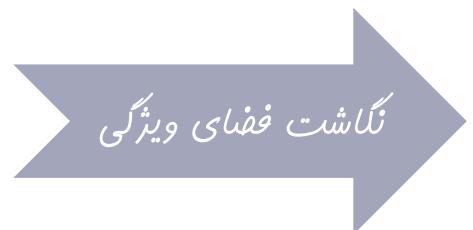
$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$

□ نقاط راهنمایی.

$$l^{(1)} = x^{(1)}, l^{(2)} = x^{(2)}, \dots, l^{(m)} = x^{(m)}$$

□ نگاشت فضای ویژگی.

$$x = \begin{bmatrix} x_0 = 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$



$$f = \begin{bmatrix} f_0 = 1 \\ f_1 = K(x, l^{(1)}) \\ f_2 = K(x, l^{(2)}) \\ \vdots \\ f_m = K(x, l^{(m)}) \end{bmatrix}$$

ترفند کرnel

۳۴

□ تابع کرnel. پیش‌پردازش داده x با استفاده از توابع کرnel:

$$\begin{aligned}\mathbf{z} &= \varphi(\mathbf{x}) \\ &= (\varphi_1(\mathbf{x}), \varphi_2(\mathbf{x}), \dots, \varphi_k(\mathbf{x}))\end{aligned}$$

$$\begin{aligned}g(\mathbf{z}) &= \mathbf{w}^T \mathbf{z} + b \\ g(\mathbf{x}) &= \mathbf{w}^T \varphi(\mathbf{x}) + b\end{aligned}$$

ممکن است بینوایت باشد!!!

□ مرز تصمیم‌گیری.

$$\mathbf{w} = \sum_{t=1}^m \alpha^t y^t \mathbf{z}^t = \sum_{t=1}^m \alpha^t y^t \varphi(\mathbf{x}^t)$$

□ دسته‌بندی داده جدید.

$$g(\mathbf{x}) = \mathbf{w}^T \varphi(\mathbf{x}) + b = \left(\sum_{t=1}^m \alpha^t y^t \varphi(\mathbf{x}^t)^T \right) \varphi(\mathbf{x}) + b = \left(\sum_{t=1}^m \alpha^t y^t \varphi(\mathbf{x}^t)^T \varphi(\mathbf{x}) \right) + b = \left(\sum_{t=1}^m \alpha^t y^t k(\mathbf{x}^t, \mathbf{x}) \right) + b$$

توابع کرnel

۳۵

$$L_p = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{t=1}^m \varepsilon^t$$

s.t. $y^t \mathbf{w}^T \varphi(\mathbf{x}^t) \geq 1 - \varepsilon^t$

$$\varepsilon^t \geq 0$$

$$L_p = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{t=1}^m \varepsilon^t - \sum_{t=1}^m \alpha^t [y^t \mathbf{w}^T \varphi(\mathbf{x}^t) - 1 + \varepsilon^t] - \sum_{t=1}^m \mu^t \varepsilon^t$$

↑
ضرایب لگرانژ ↑
ضرایب لگرانژ

توابع کرnel: مسئله اصلی

۳۶

$$L_p = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{t=1}^m \varepsilon^t - \sum_{t=1}^m \alpha^t [y^t \mathbf{w}^T \varphi(\mathbf{x}^t) - 1 + \varepsilon^t] - \sum_{t=1}^m \mu^t \varepsilon^t$$

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{t=1}^m \alpha^t y^t \varphi(\mathbf{x}^t)$$

$$\frac{\partial L_p}{\partial \varepsilon^t} = 0 \Rightarrow C - \alpha^t - \mu^t = 0 \Rightarrow 0 \leq \alpha^t \leq C$$

توابع کرnel: مسئله دوگان

۳۷

$$L_d = -\frac{1}{2} \sum_{t=1}^m \sum_{s=1}^m \alpha^t \alpha^s y^t y^s \varphi(\mathbf{x}^t)^T \varphi(\mathbf{x}^s) + \sum_{t=1}^m \alpha^t$$

subject to $\sum_{t=1}^m \alpha^t y^t = 0$ and $0 \leq \alpha^t \leq C \forall t$

□ ایده ماشین‌های کرnel. [ترفند کرnel]

□ جایگزینی حاصل‌ضرب داخلی توابع پایه با یک تابع کرnel به صورت $K(\mathbf{x}^t, \mathbf{x}^s)$

$$L_d = -\frac{1}{2} \sum_{t=1}^m \sum_{s=1}^m \alpha^t \alpha^s y^t y^s K(\mathbf{x}^t, \mathbf{x}^s) + \sum_{t=1}^m \alpha^t$$

ماتریس K : یک ماتریس متقارن و مثبت معین (برای تفکیک پذیری فطی)

توابع کرnel: کرnel چندجمله‌ای

۳۸

کرnel چند جمله‌ای. یک چندجمله‌ای از درجه q . □

$$K(x^t, x) = (x^T x^t + 1)^q$$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

[$q = 2, d = 2$] مثال. □

$$K(x, y) = (x^T y + 1)^2$$

۳ ضرب، ۲ جمع

$$= (x_1 y_1 + x_2 y_2 + 1)^2$$

۶ ضرب، ۵ جمع

$$= 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 x_2 y_1 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2$$

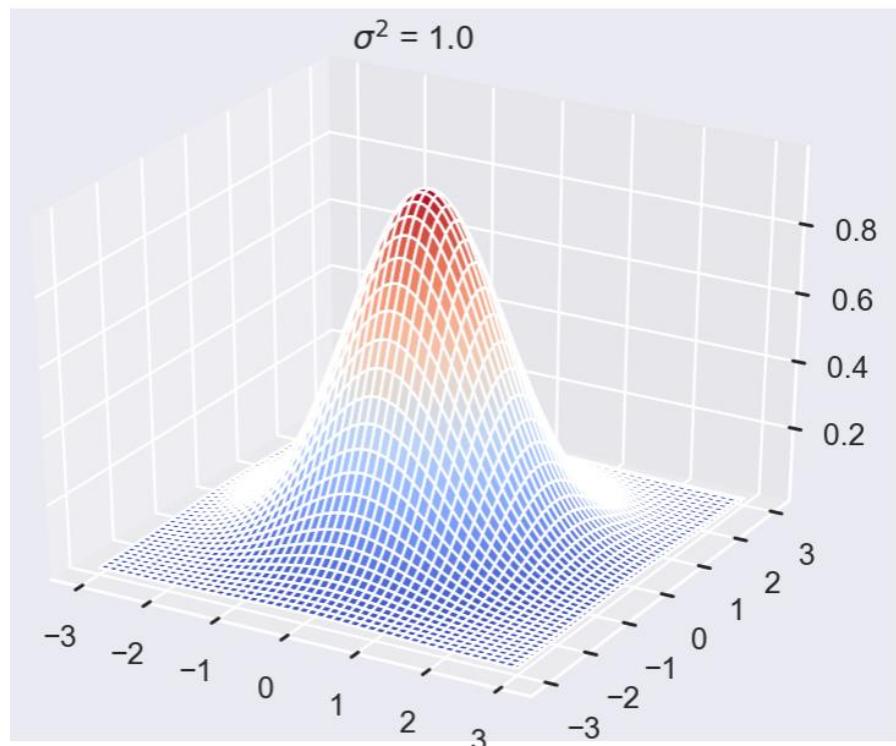
$$\varphi(x) = [1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2]^T$$

$$\varphi(y) = [1, \sqrt{2}y_1, \sqrt{2}y_2, \sqrt{2}y_1 y_2, y_1^2, y_2^2]^T$$

توابع کرnel: کرnel گوسی

۳۹

$$K(x^t, x) = \exp\left(-\frac{\|x^t - x\|^2}{2\sigma^2}\right)$$

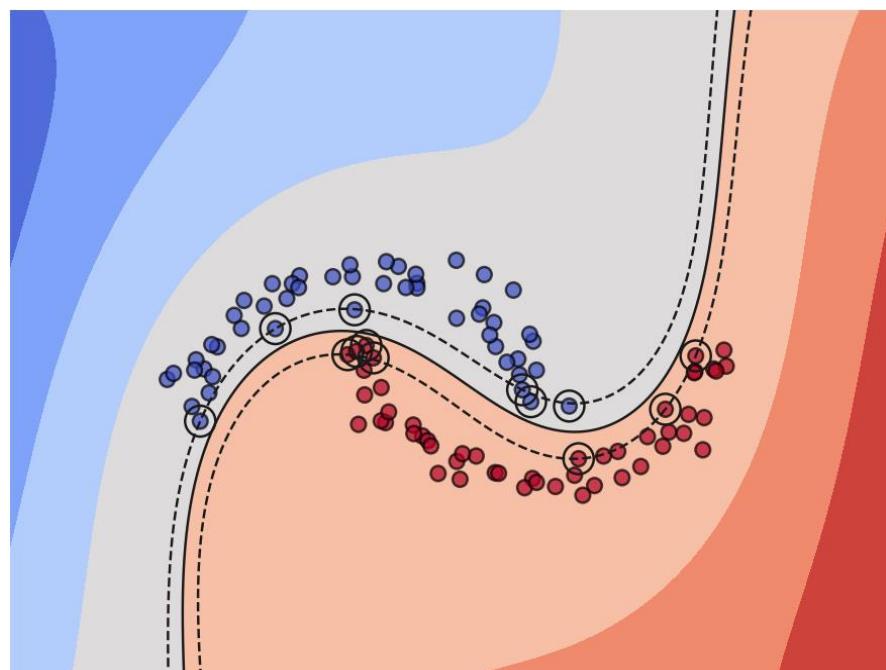


- کرnel گوسی.
- یافتن یک مقدار مناسب برای σ .
- با استفاده از مجموعه اعتبارسنجی [انتخاب مدل]
- مقادیر بزرگتر: مرز تصمیم‌گیری هموارتر

توابع کرnel: کرnel گوسی

۴۰

$$K(x^t, x) = \exp\left(-\frac{\|x^t - x\|^2}{2\sigma^2}\right)$$



- کرnel گوسی.
- یافتن یک مقدار مناسب برای σ .
- با استفاده از مجموعه اعتبارسنجی [انتخاب مدل]
- مقادیر بزرگتر: مرز تصمیم‌گیری هموارتر

ابر پارامترها: انتخاب مدل

۴۱

- راه حل اول. [یک ایده بسیار بد]
 - انتخاب مقداری که منجر به بالاترین دقت دسته‌بندی بر روی **مجموعه آزمایشی** می‌شود.

داده‌های آموزشی	داده‌های آزمایشی
-----------------	------------------

- توجه. [بسیار مهم]
 - از مجموعه آزمایشی در انتهای مراحل و تنها برای **تخمین قابلیت تعمیم** دسته‌بند استفاده کنید.

ابر پارامترها: انتخاب مدل

۴۲

□ راه حل دوم. [اعتبارسنجی چند بخشی]

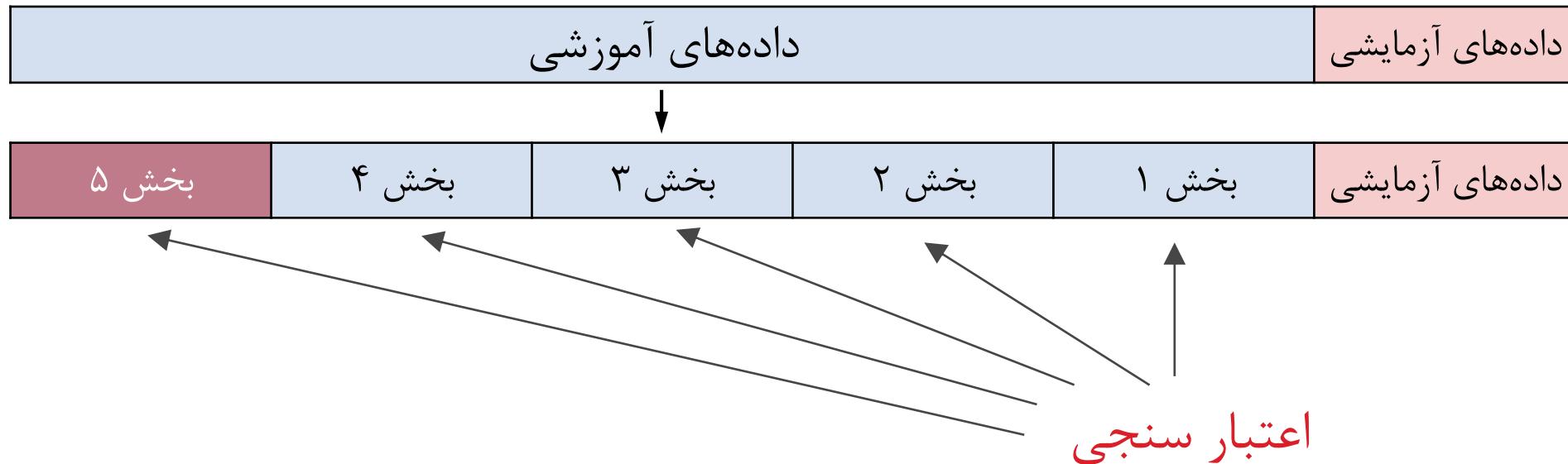
داده‌های آموزشی					داده‌های آزمایشی
بخش ۵	بخش ۴	بخش ۳	بخش ۲	بخش ۱	داده‌های آزمایشی
بخش ۵	بخش ۴	بخش ۳	بخش ۲	بخش ۱	داده‌های آزمایشی
بخش ۵	بخش ۴	بخش ۳	بخش ۲	بخش ۱	داده‌های آزمایشی
بخش ۵	بخش ۴	بخش ۳	بخش ۲	بخش ۱	داده‌های آزمایشی
بخش ۵	بخش ۴	بخش ۳	بخش ۲	بخش ۱	داده‌های آزمایشی

داده‌های اعتبارسنجی [برای تعیین مقدار ابرپارامترها]

ابر پارامترها: انتخاب مدل

۴۳

□ راه حل دوم. [اعتبارسنجی چند بخشی]



هر بار یک بخش را به عنوان داده‌های اعتبارسنجی
انتخاب کن و سپس از نتایج به دست آمده میانگین بگیر

ابر پارامترها: انتخاب مدل

۴۴

```
cv = StratifiedKFold(n_splits=5, shuffle=True)
```

تقسیم داده‌های آموزشی
 به داده‌های آموزشی و
 داده‌های اعتبارسنجی

```
C_range = np.logspace(-3, 5, 9)
```

```
gamma_range = np.logspace(-3, 5, 9)
```

```
pgrid = dict(gamma=gamma_range, C=C_range)
```

```
grid = GridSearchCV(SVC(), param_grid=pgrid, cv=cv)
```

```
grid.fit(X, y)
```

ابر پارامترها: انتخاب مدل

۴۵

```
cv = StratifiedKFold(n_splits=5, shuffle=True)
```

```
C_range = np.logspace(-3, 5, 9)  
gamma_range = np.logspace(-3, 5, 9)
```

مشخص کردن بازه
جستجو برای ابرپارامترها

```
pgrid = dict(gamma=gamma_range, C=C_range)  
grid = GridSearchCV(SVC(), param_grid=pgrid, cv=cv)
```

```
grid.fit(X, y)
```

$$\{10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4, 10^5\}$$

ابر پارامترها: انتخاب مدل

۴۶

```
cv = StratifiedKFold(n_splits=5, shuffle=True)

C_range = np.logspace(-3, 5, 9)
gamma_range = np.logspace(-3, 5, 9)

pgrid = dict(gamma=gamma_range, C=C_range)
grid = GridSearchCV(SVC(), param_grid=pgrid, cv=cv)

grid.fit(X, y)
```

ایجاد دسته‌بند

ابر پارامترها: انتخاب مدل

۴۷

```
cv = StratifiedKFold(n_splits=5, shuffle=True)

C_range = np.logspace(-3, 5, 9)
gamma_range = np.logspace(-3, 5, 9)

pgrid = dict(gamma=gamma_range, C=C_range)
grid = GridSearchCV(SVC(), param_grid=pgrid, cv=cv)

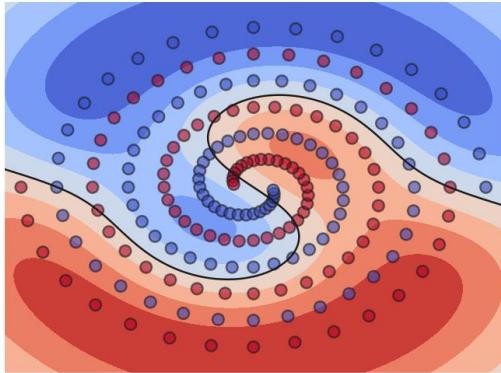
grid.fit(X, y)
```

آموزش دسته‌بند

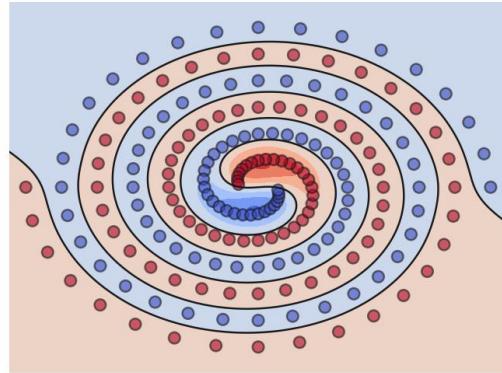
جسنجوی تواری

۴۸

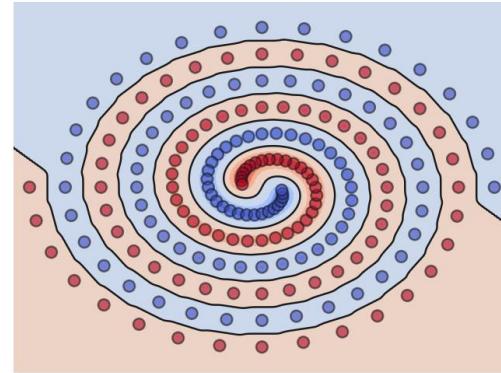
$$\gamma = 10^{-1}, C = 10^{-1}$$



$$\gamma = 10^0, C = 10^{-1}$$

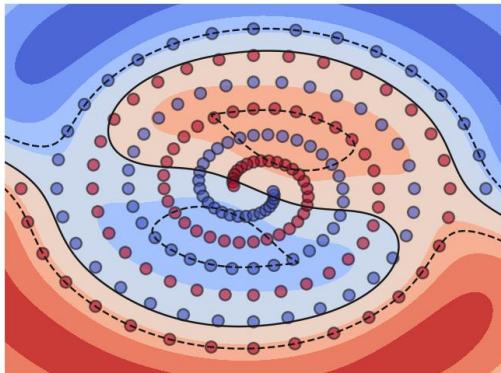


$$\gamma = 10^1, C = 10^{-1}$$

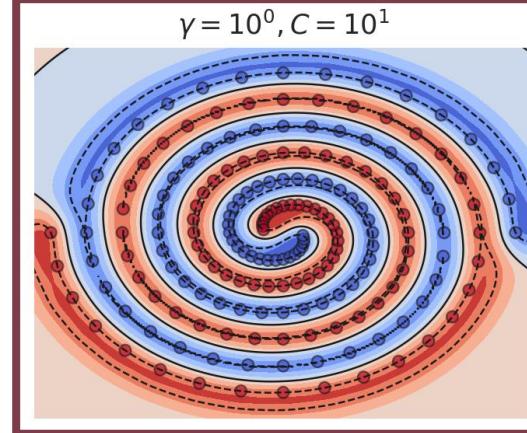


$$C = 10^{-1}$$

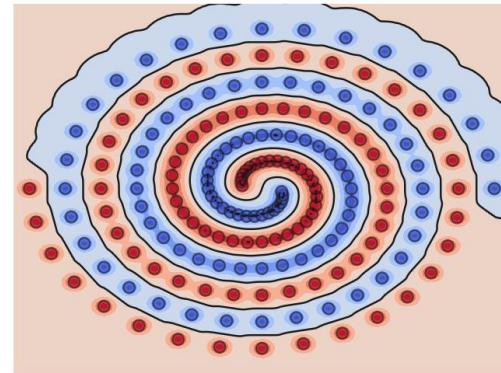
$$\gamma = 10^{-1}, C = 10^1$$



$$\gamma = 10^0, C = 10^1$$

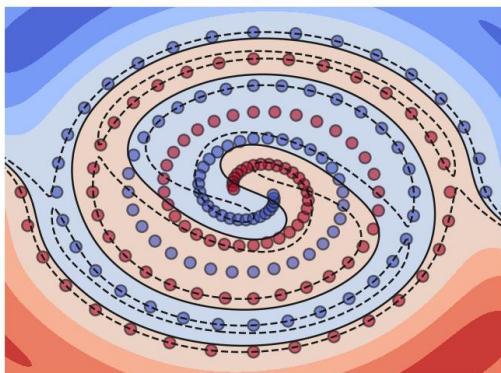


$$\gamma = 10^1, C = 10^1$$

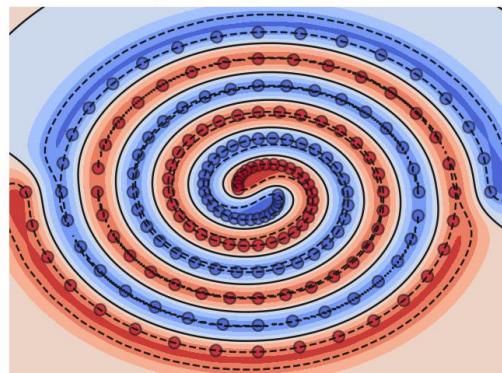


$$C = 10^1$$

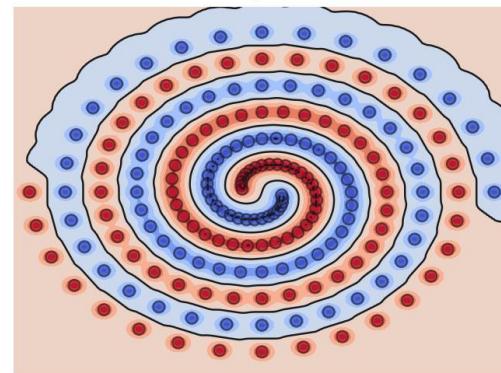
$$\gamma = 10^{-1}, C = 10^3$$



$$\gamma = 10^0, C = 10^3$$



$$\gamma = 10^1, C = 10^3$$



$$C = 10^3$$

$$\gamma = 10^{-1}$$

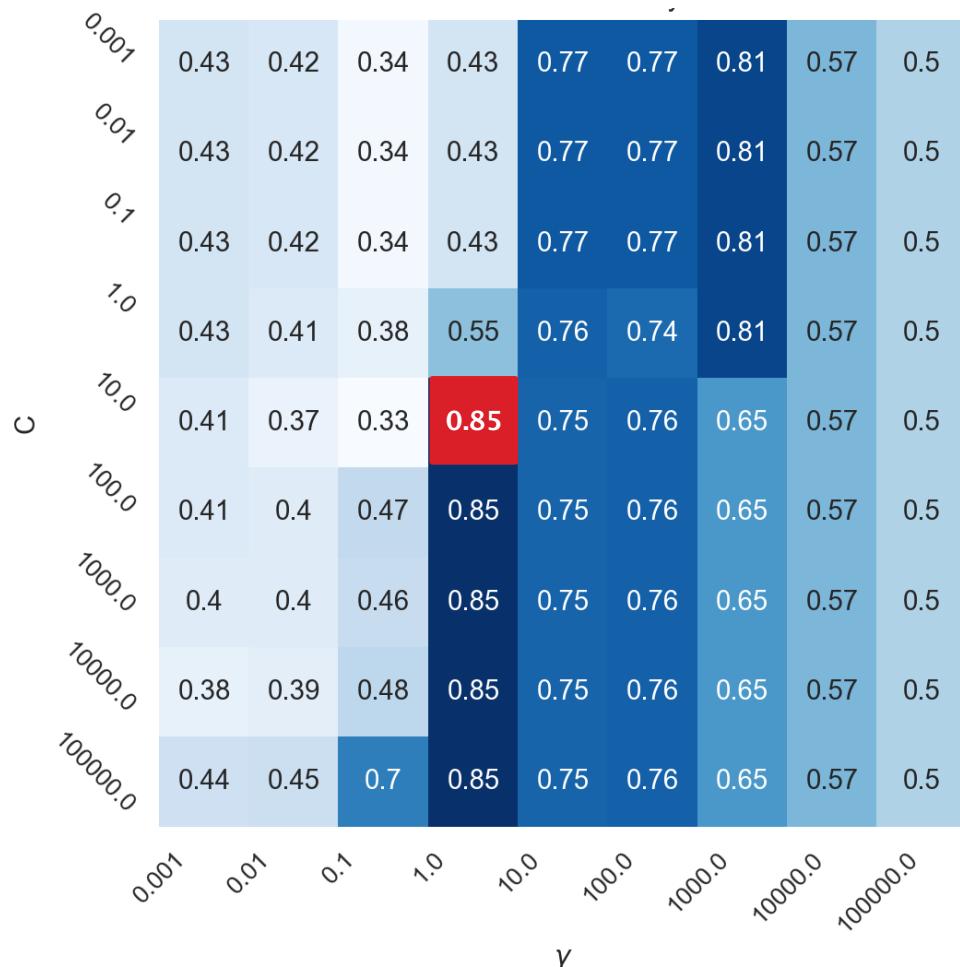
$$\gamma = 10^0$$

$$\gamma = 10^1$$

پارامترهای ماشین بردار پشتیبان

۴۹

□ پرسش: مقدار مناسب برای پارامترهای تابع کرنل چگونه تعیین می‌شوند؟



دقت دسته‌بندی بر روی داده‌های اعتبارسنجی

□ پارامتر C

■ مقادیر کوچک‌تر: بایاس بیشتر، واریانس کمتر

■ مقادیر بزرگ‌تر: بایاس کمتر، واریانس بیشتر

□ پارامتر σ

■ مقادیر کوچک‌تر: بایاس کمتر، واریانس بیشتر

■ مقادیر بزرگ‌تر: بایاس بیشتر، واریانس کمتر

راهنمای استفاده از ماشین بردار پشتیبان

۵۰

- پیاده‌سازی. استفاده از بسته‌های نرم‌افزاری موجود مانند LIBSVM^{light} و SVM^{light}
- تعیین تابع کرنل.
- کرنل خطی (عدم استفاده از کرنل): وقتی که $n \gg m$
- گوسی، چندجمله‌ای، رشته‌ای و ...
- تعیین مقدار پارامترها. جستجوی توری
- انتخاب مقدار برای پارامتر C
- انتخاب مقدار برای پارامترهای تابع کرنل (مانند σ)

ماشین بردار پشتیبان، شبکه عصبی یا رگرسیون لجستیک

۵۱

□ حالت ۱. $[n \gg m]$

- مثال: تشخیص هرزname (۱۰۰۰ نمونه آموزشی، ۵۰۰۰۰ ویژگی)
- رگرسیون لجستیک یا ماشین بردار پشتیبان خطی

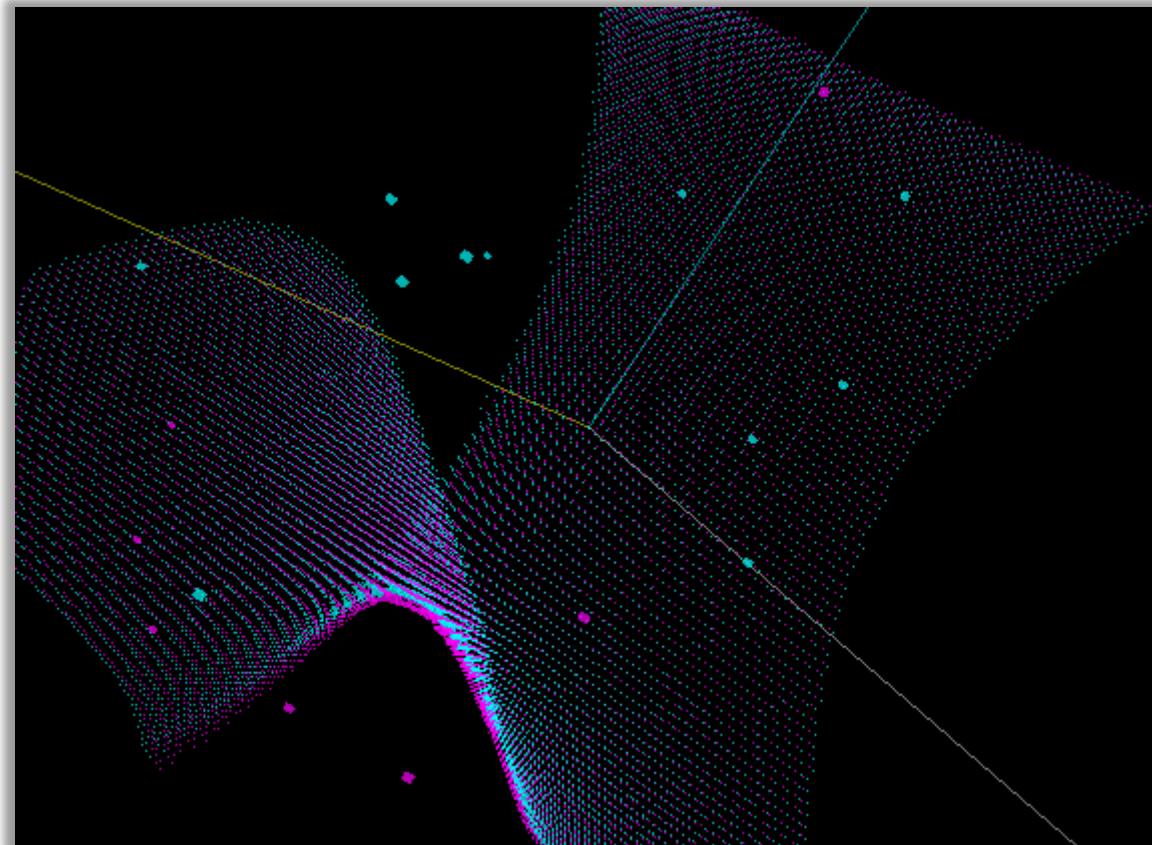
□ حالت ۲. [تعداد ویژگی‌ها کم، تعداد نمونه‌های آموزشی زیاد]

- ماشین بردار پشتیبان با کرنل گوسی

□ **توجه.** شبکه‌های عصبی در تمامی حالت‌های فوق قابل استفاده هستند، اما ممکن است به زمان بیشتری برای آموزش نیاز داشته باشند.

اجرای نمایشی

۵۲



<http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/svmtoy3d/>

پیوست: مطالب بیشتر در مورد کرnelها

۵۳

- پرسش. چگونه می‌دانیم استفاده از کرnelها در جداسازی داده‌ها به ما کمک می‌کند؟
 - در فضای n -بعدی، هر مجموعه از n بردار مستقل، به صورت خطی تفکیک‌پذیر هستند.
 - اگر ماتریس K یک ماتریس مثبت معین باشد، آنگاه داده‌ها به صورت خطی تفکیک‌پذیر هستند.

- قضیه. ماتریس K یک ماتریس مثبت معین است، زیرا $K = L^T L$
 - ستون i در ماتریس L برابر است با بردار $\phi(x^{(i)})$

- اثبات. بردار غیر صفر v را در نظر بگیرید. در این صورت:
$$v^T K v = v^T L^T L v = (Lv)^T (Lv) = w^T w = \|w\|^2 \geq 0$$
و چون L و v هر دو مخالف صفر هستند، بردار w نیز مخالف صفر است. یعنی:
$$\|w\|^2 > 0 \Rightarrow v^T K v > 0 \Rightarrow K \text{ is positive definite}$$

یادگیری بدون نظارت: فوشنگندی

سید ناصر رضوی www.snrazavi.ir

۱۳۹۷

فهرست مطالب

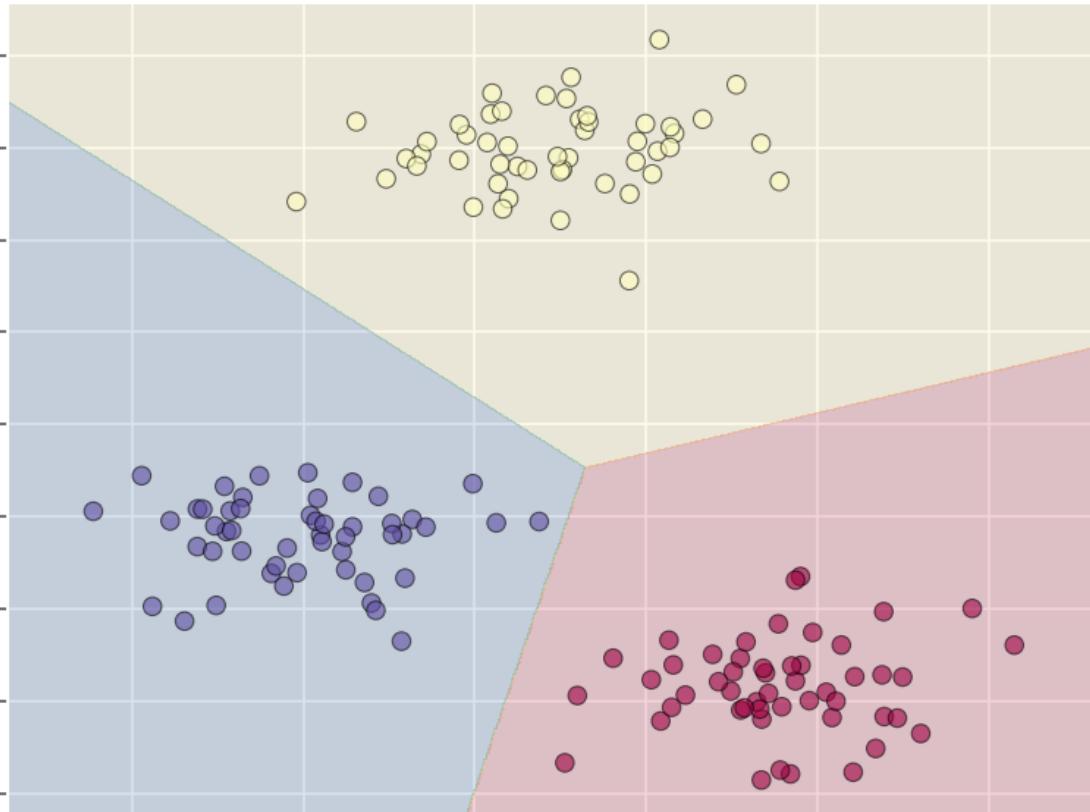
۲

- یادگیری بدون نظارت
- خوشه‌بندی
- الگوریتم K-means
- بهبود خوشه‌بندی
- الگوریتم دو بخشی‌ساز
- خوشه‌بندی سلسله‌مراتبی

یادآوری: یادگیری نظارت شده

۳

- یادگیری نظارت شده. به ازای هر نمونه، پاسخ درست داده شده است.



$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$

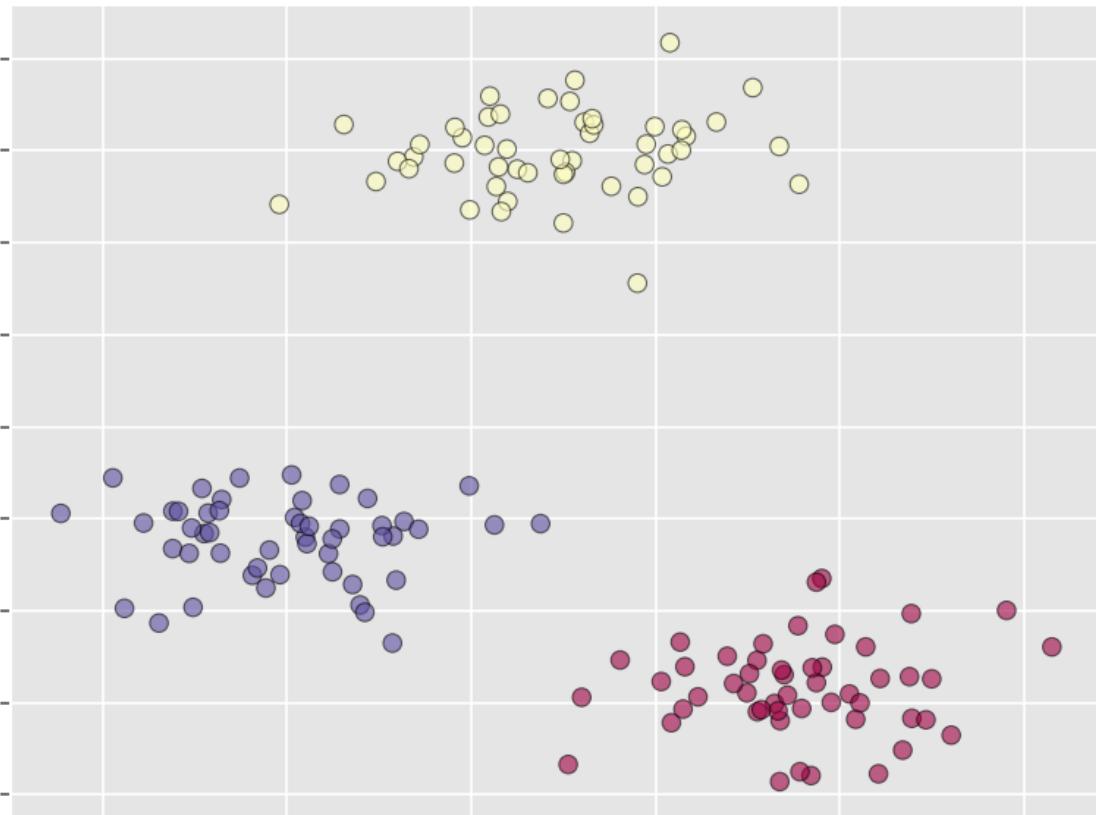
مجموعه آموزشی

- انواع یادگیری نظارت شده.
 - رگرسیون: تخمین یک کمیت پیوسته
 - دسته‌بندی: تخمین یک کمیت گستته

یادگیری بدون نظارت

۴

□ یادگیری بدون نظارت. عدم آگاهی از پاسخ‌های درست.



$$\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(m)}\}$$

مجموعه آموزشی

□ هدف. تشخیص ساختار در داده‌های ورودی

کاربرد خوشنودی افبار مرتبه

The image displays four web browser windows side-by-side, each showing a different news source reporting on the Deepwater Horizon oil spill.

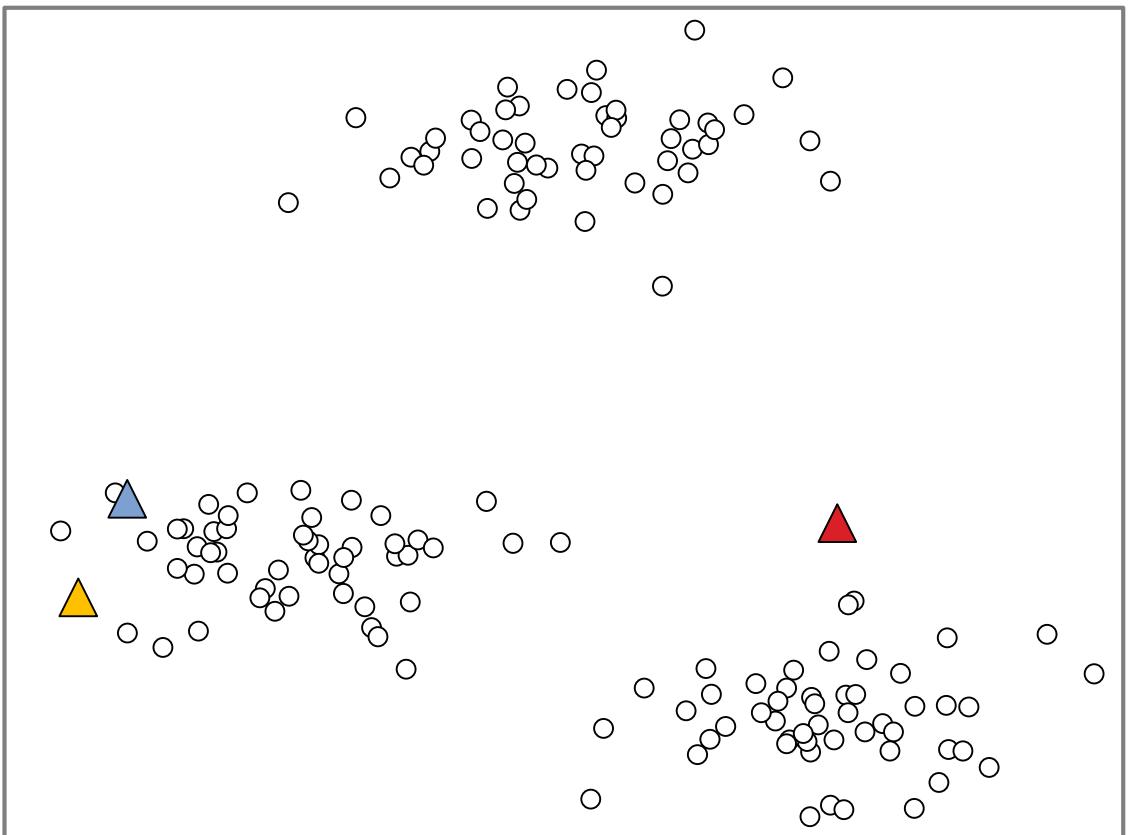
- Google News (Left):** Shows a list of top stories, including "BP Oil Well, Site of National Catastrophe, Dies at One" and "Allen: Well is dead, but much Gulf Coast work remains".
- CNN (Second from Left):** Shows the headline "Allen: Well is dead, but much Gulf Coast work remains" with a video player showing an oil rig.
- THE WALL STREET JOURNAL SOURCE (Second from Right):** Shows the headline "BP Kills Macondo, But Its Legacy Lives On" with a large image of a burning oil platform.
- guardian.co.uk (Right):** Shows the headline "BP oil spill cost hits nearly \$10bn" with a large image of a burning oil platform.

Red arrows indicate the flow of information or the interconnected nature of these news stories:

- An arrow points from the Google News screenshot to the CNN screenshot.
- An arrow points from the CNN screenshot to the WSJ Source screenshot.
- An arrow points from the WSJ Source screenshot to the guardian.co.uk screenshot.

الگوریتم خوشه‌بندی K-means

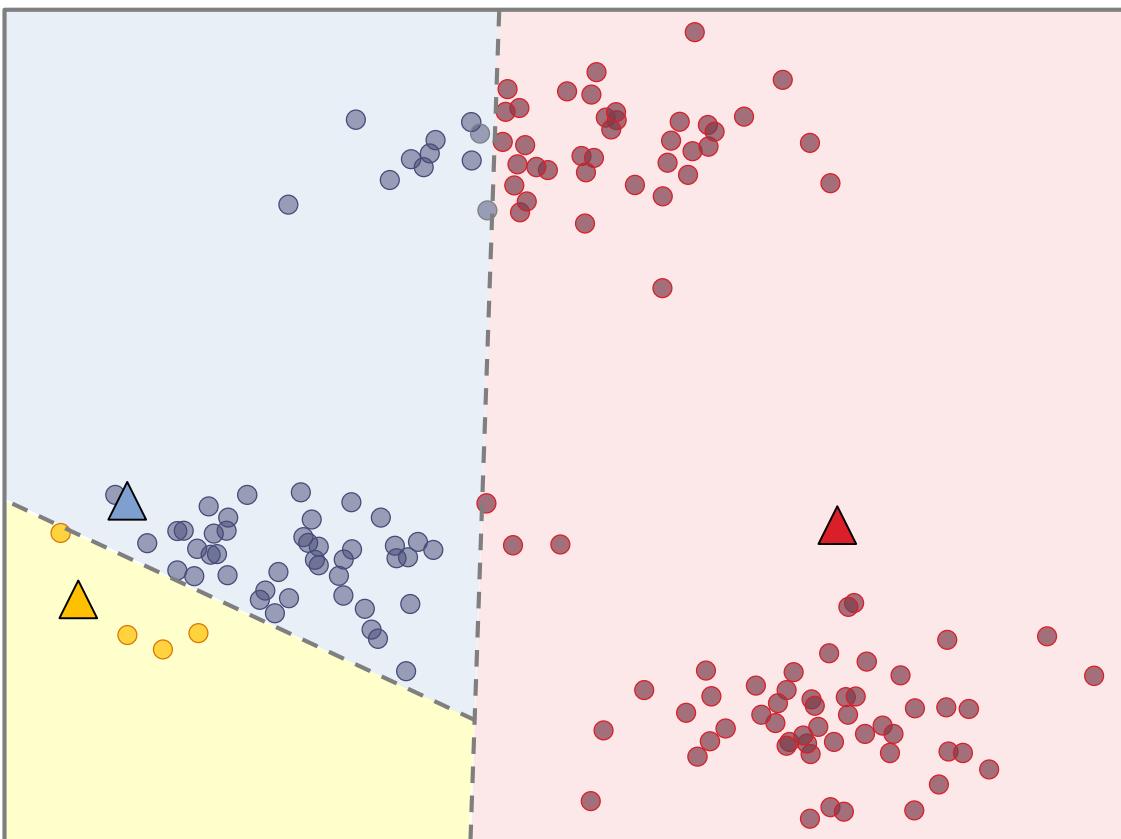
۶



- یک الگوریتم خوشه‌بندی تکرار شونده.
- K نقطه را به صورت تصادفی به عنوان مراکز خوشه‌ها انتخاب کن.
- مراحل زیر را تکرار کن:
 - هر داده را به یک خوشه با نزدیکترین مرکز انتساب بده.
 - مرکز هر خوشه را با میانگین‌گیری از داده‌های انتساب یافته به آن خوشه، به روز رسانی کن.
- توقف: زمانی که در یک تکرار هیچ داده‌ای خوشه خود را عوض نکند.

الگوریتم خوشه‌بندی K-means

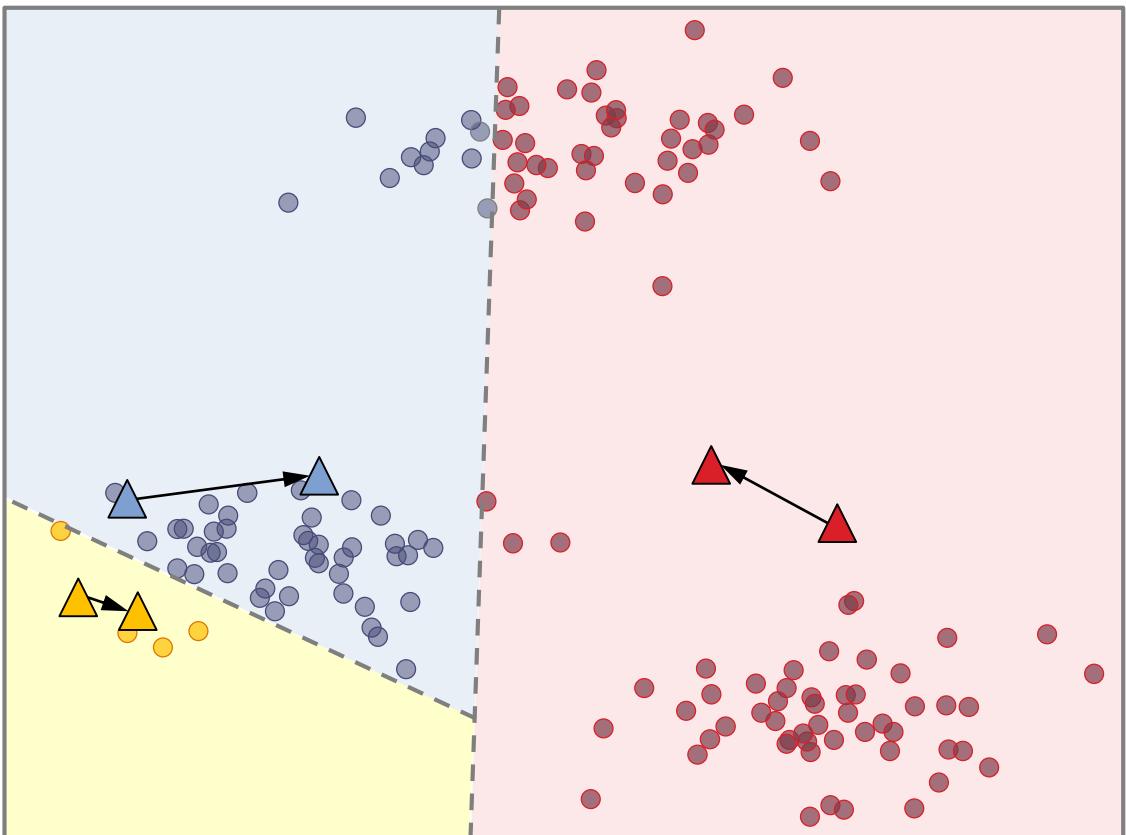
۷



- یک الگوریتم خوشه‌بندی تکرار شونده.
- K نقطه را به صورت تصادفی به عنوان مراکز خوشه‌ها انتخاب کن.
- مراحل زیر را تکرار کن:
 - هر داده را به یک خوشه با نزدیکترین مرکز انتساب بده.
 - مرکز هر خوشه را با میانگین‌گیری از داده‌های انتساب یافته به آن خوشه، به روز رسانی کن.
- توقف: زمانی که در یک تکرار هیچ داده‌ای خوشه خود را عوض نکند.

الگوریتم خوشه‌بندی K-means

۸

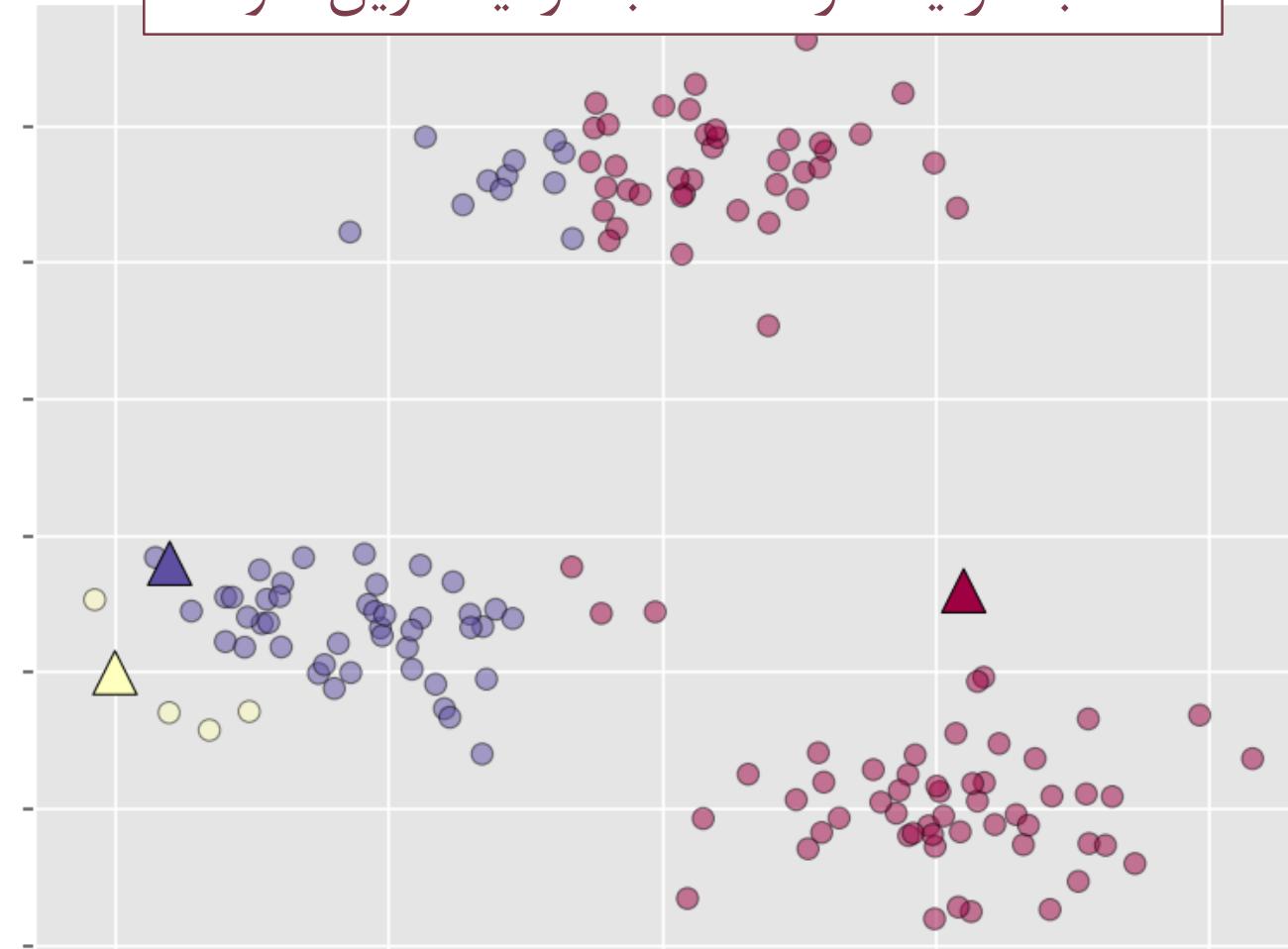


- یک الگوریتم خوشه‌بندی تکرار شونده.
- K نقطه را به صورت تصادفی به عنوان مراکز خوشه‌ها انتخاب کن.
- مراحل زیر را تکرار کن:
 - هر داده را به یک خوشه با نزدیکترین مرکز انتساب بده.
 - مرکز هر خوشه را با میانگین‌گیری از داده‌های انتساب یافته به آن خوشه، به روز رسانی کن.
- توقف: زمانی که در یک تکرار هیچ داده‌ای خوشه خود را عوض نکند.

خوشه‌بندی: اجرای نمایشی

۹

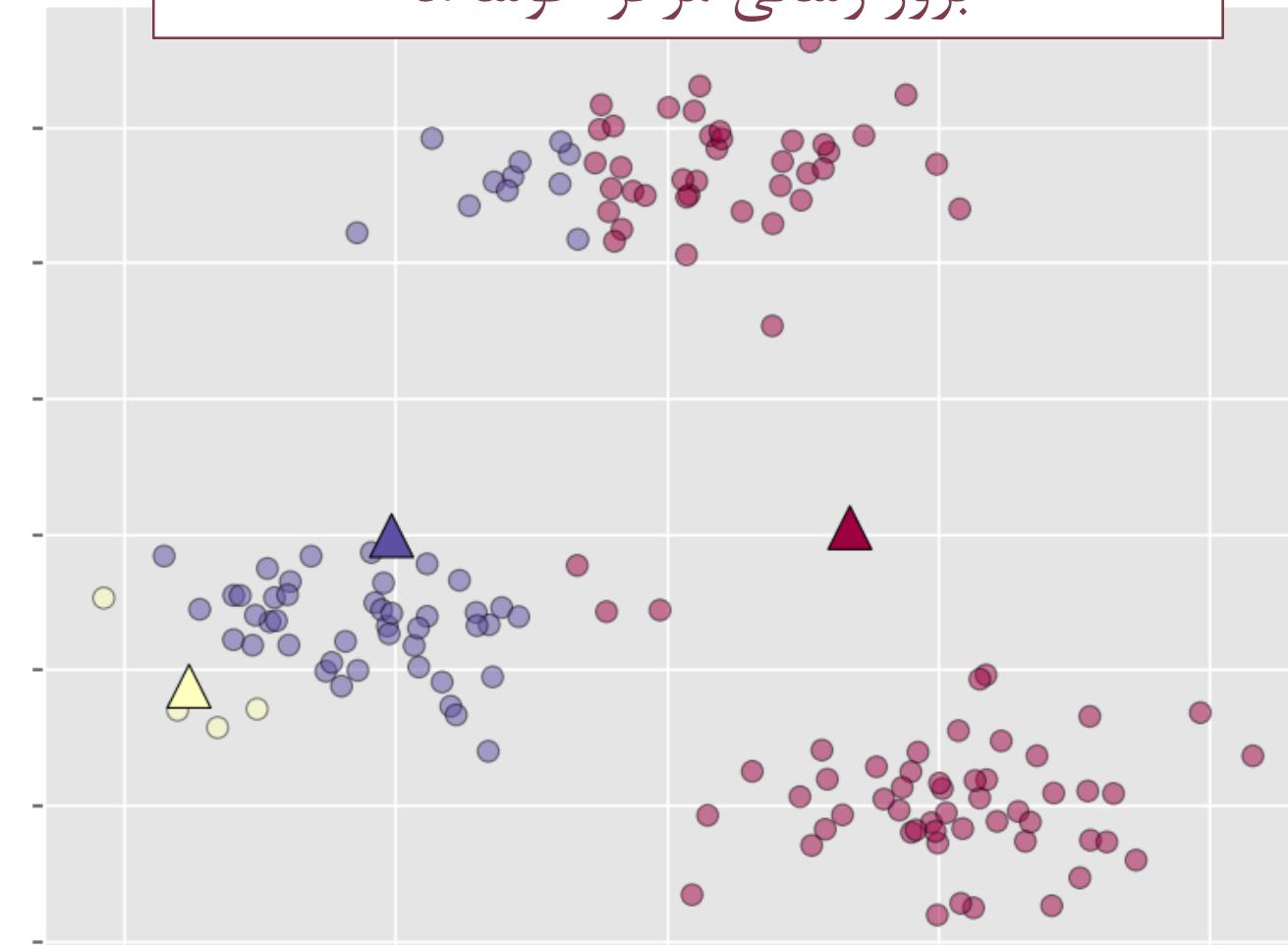
انتساب هر یک از داده‌ها به نزدیک‌ترین خوش



خوشه‌بندی: اجرای نمایشی

۱۰

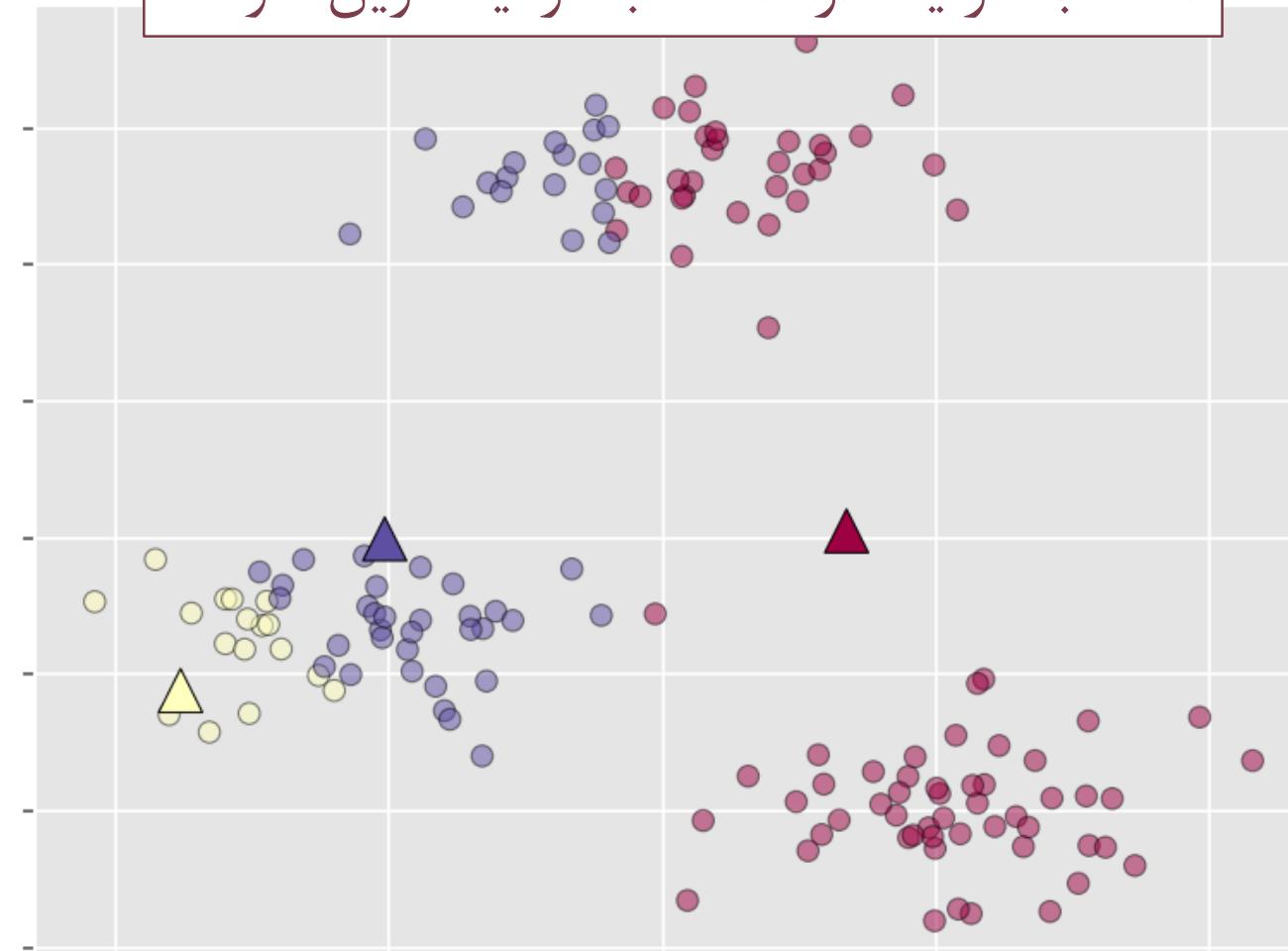
بروز رسانی مرکز خوشها



خوشه‌بندی: اجرای نمایشی

۱۱

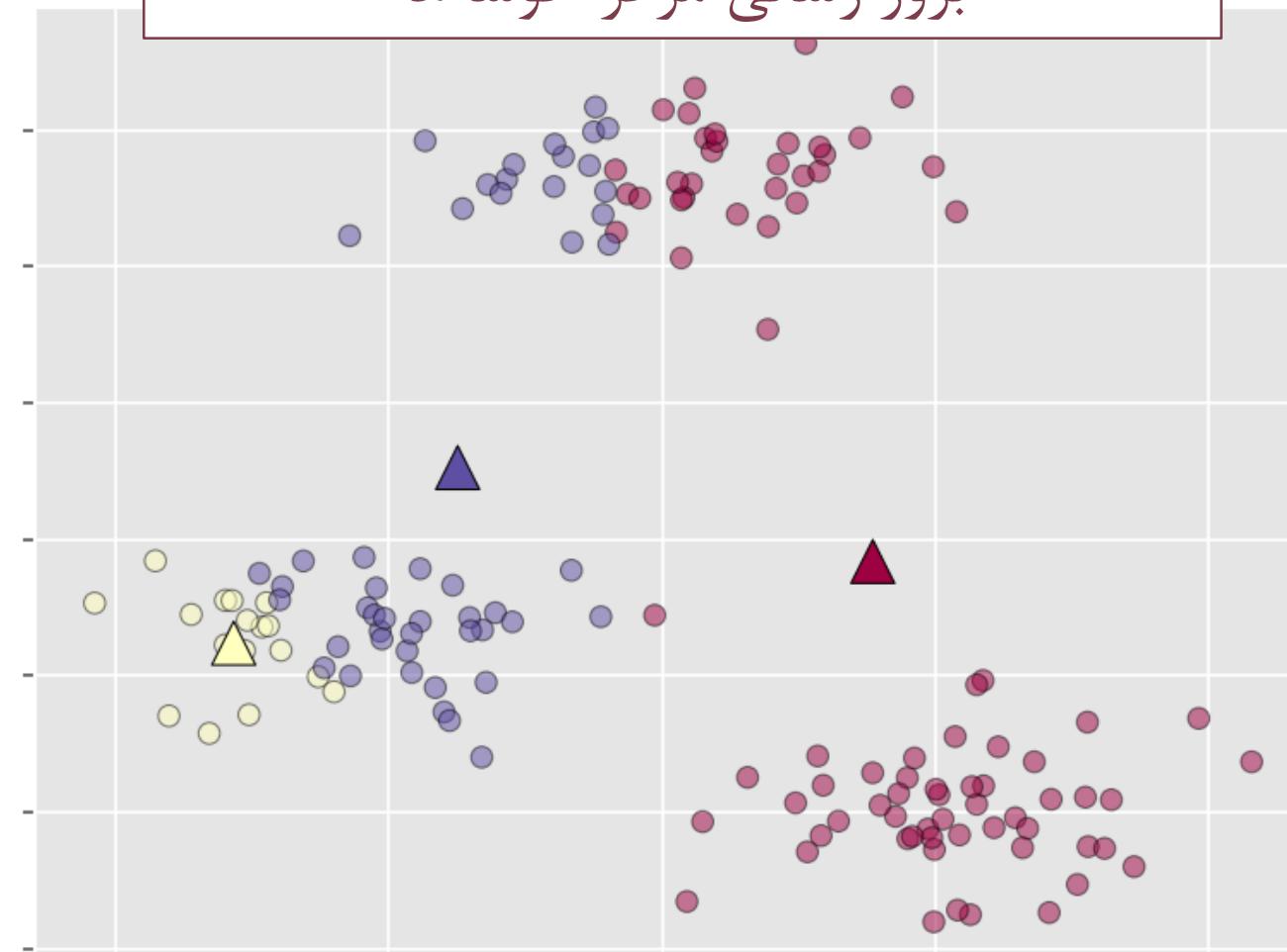
انتساب هر یک از داده‌ها به نزدیک‌ترین خوش



خوشه‌بندی: اجرای نمایشی

۱۲

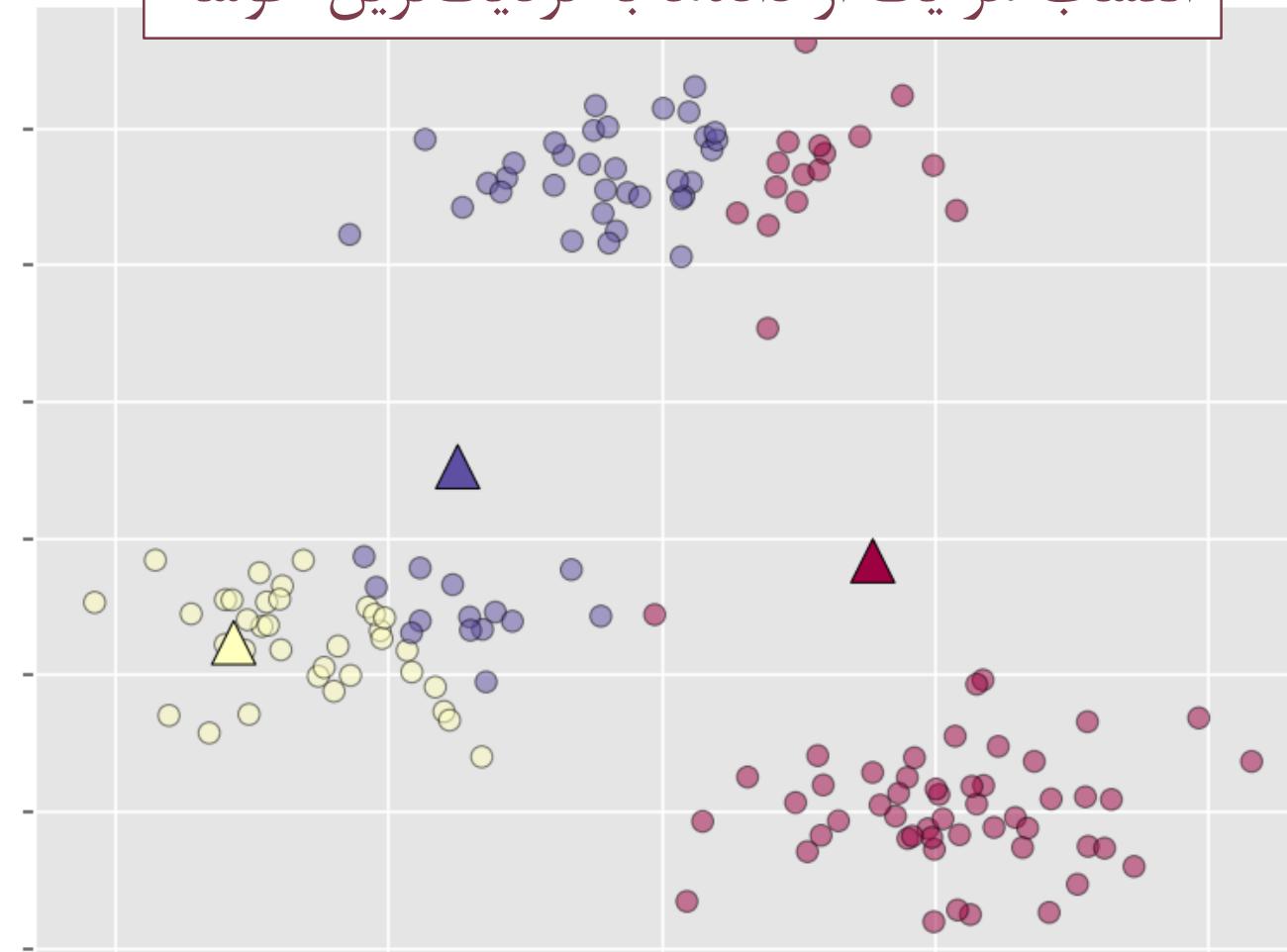
بروز رسانی مرکز خوشها



خوشه‌بندی: اجرای نمایشی

۱۳

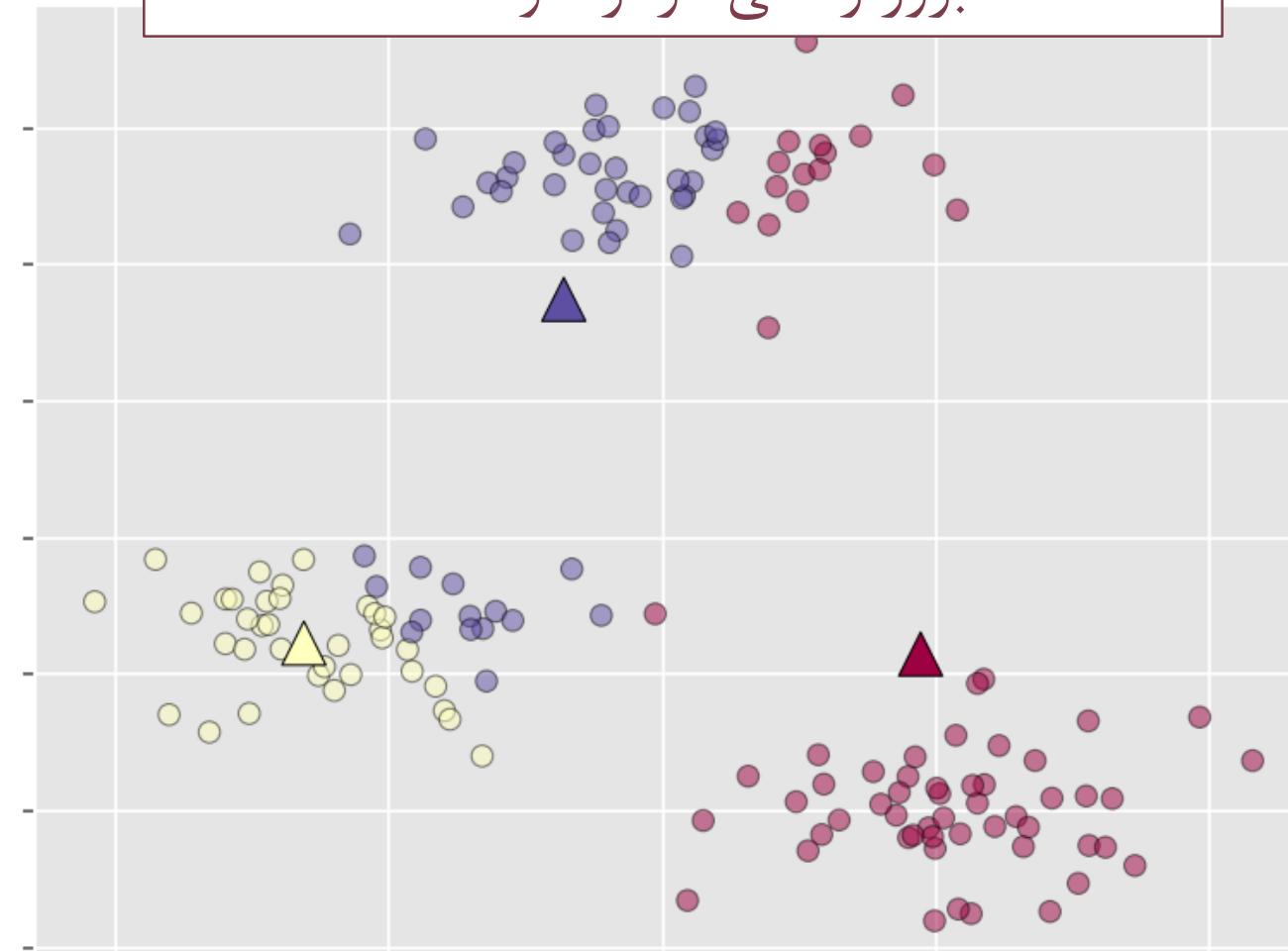
انتساب هر یک از داده‌ها به نزدیک‌ترین خوش



خوشه‌بندی: اجرای نمایشی

۱۴

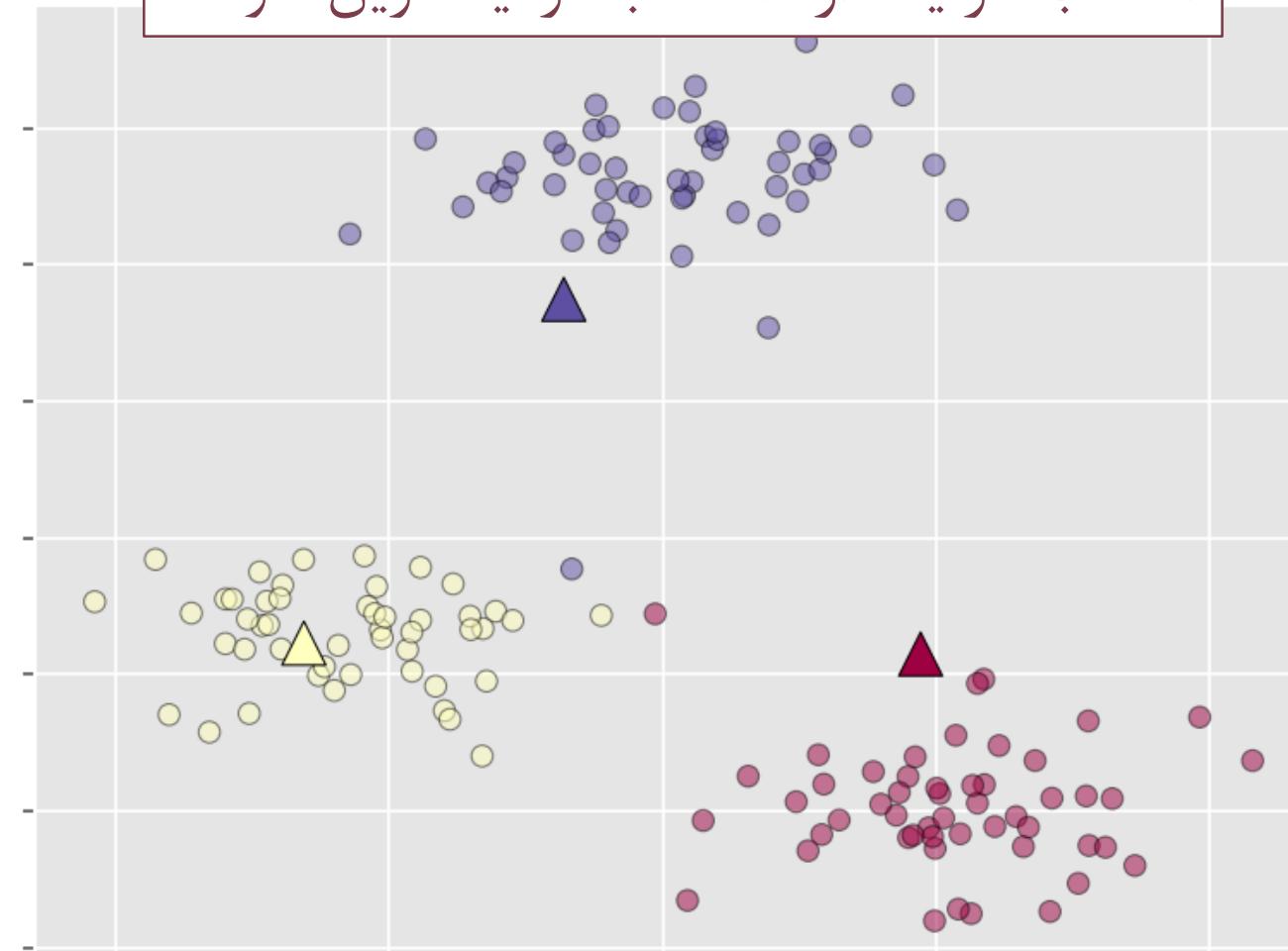
بروز رسانی مرکز خوشها



خوشه‌بندی: اجرای نمایشی

۱۵

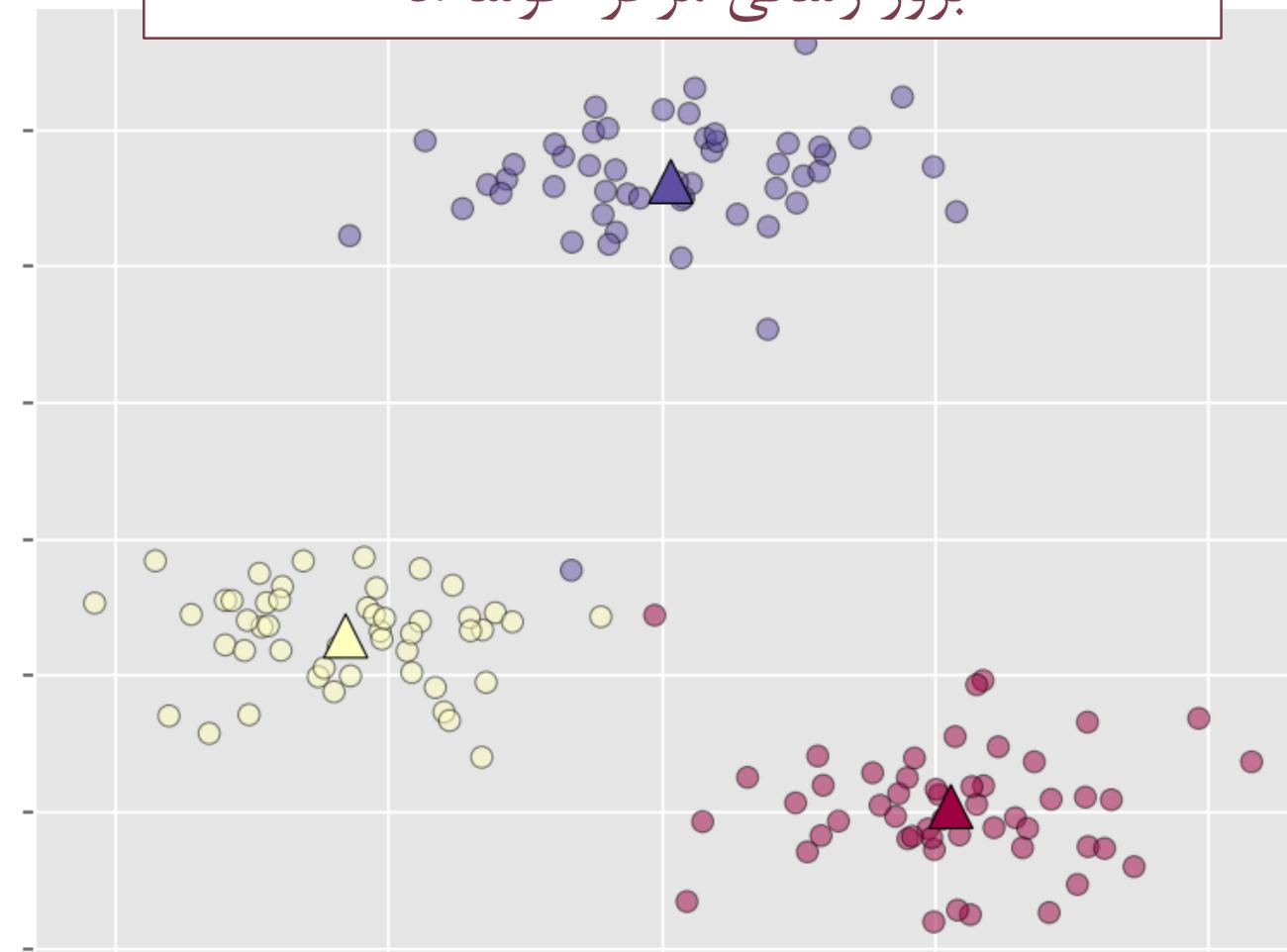
انتساب هر یک از داده‌ها به نزدیک‌ترین خوش



خوشه‌بندی: اجرای نمایشی

۱۶

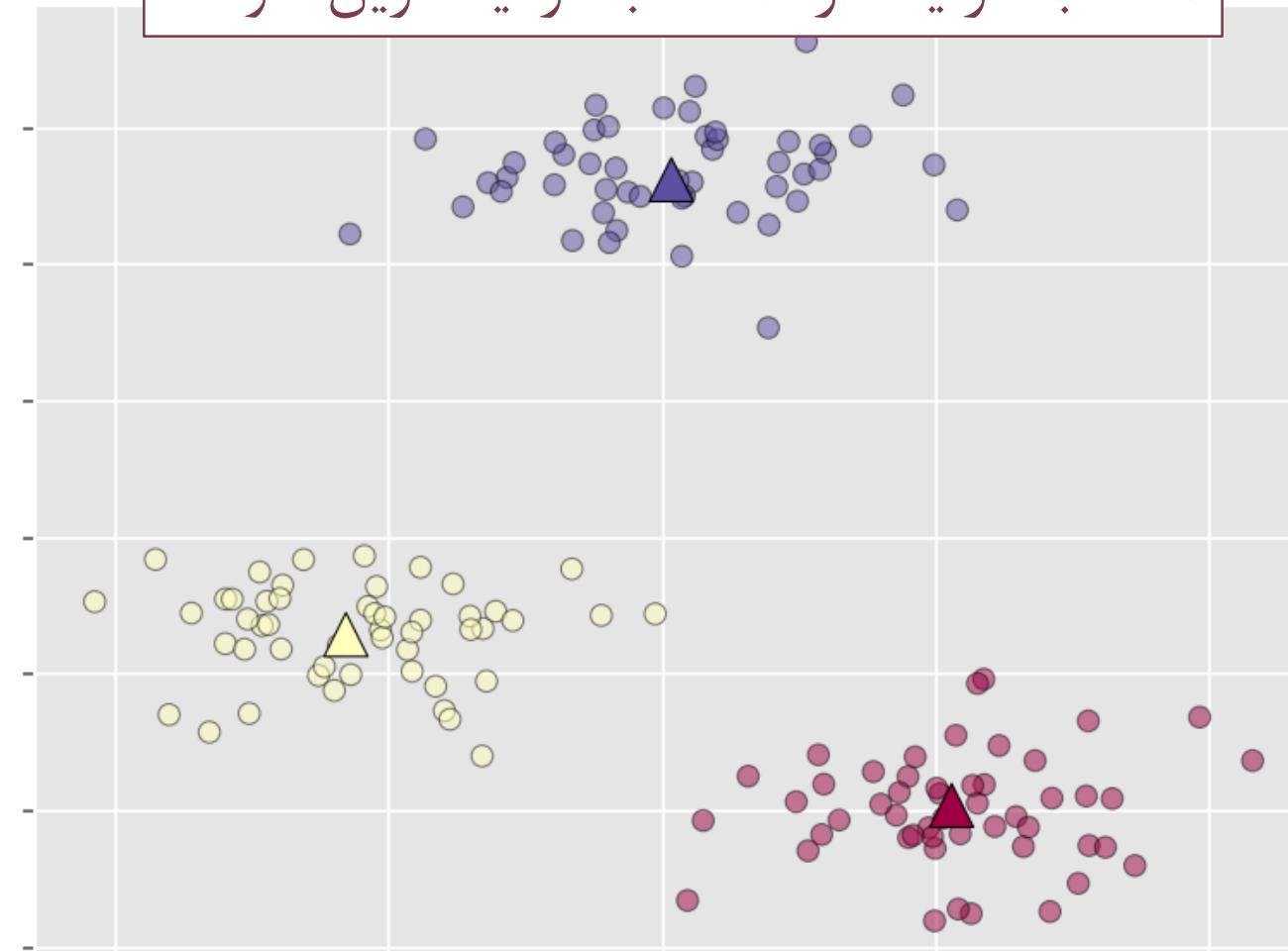
بروز رسانی مرکز خوشه‌ها



خوشه‌بندی: اجرای نمایشی

۱۷

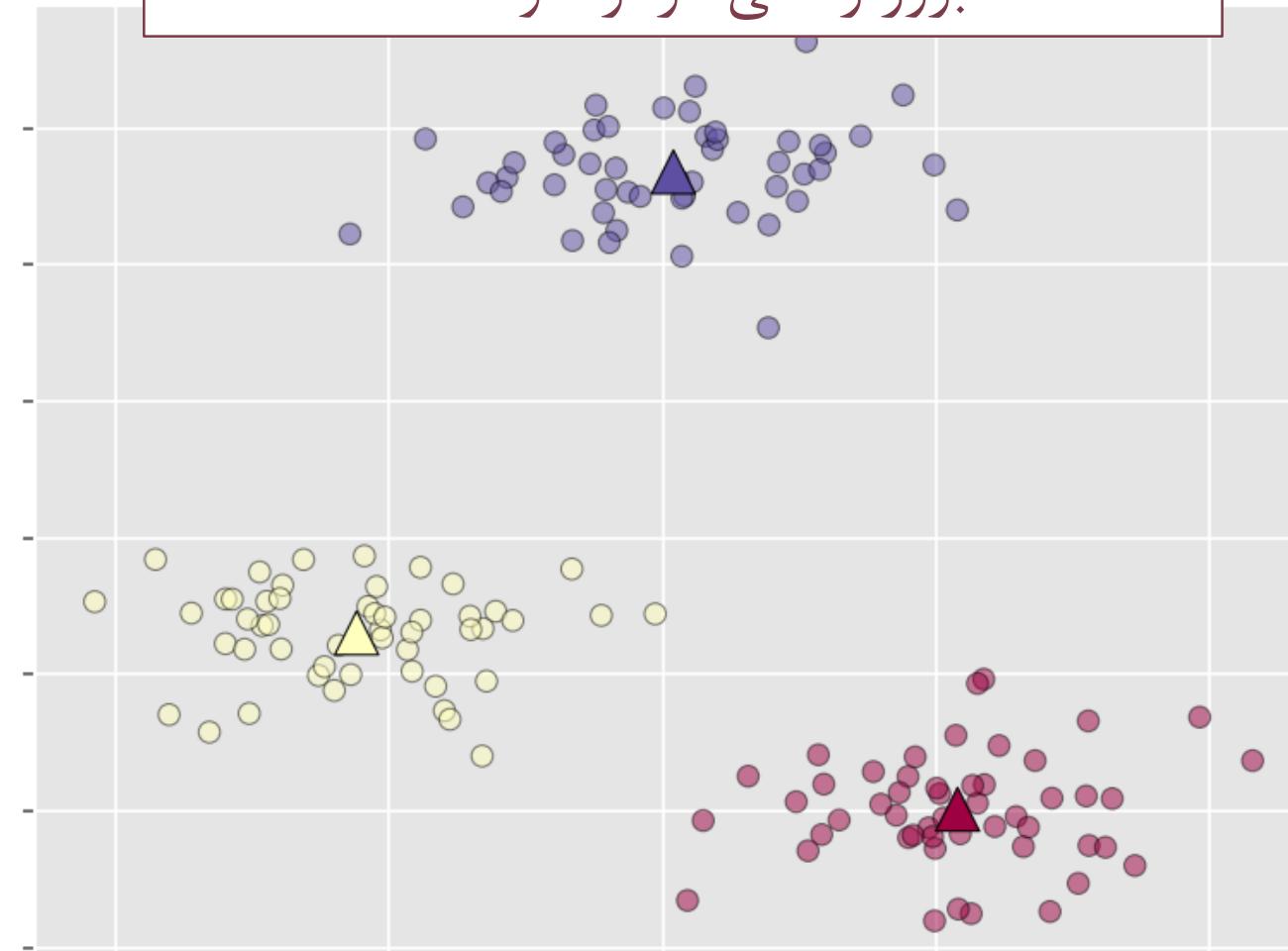
انتساب هر یک از داده‌ها به نزدیک‌ترین خوش



خوشه‌بندی: اجرای نمایشی

۱۸

بروز رسانی مرکز خوشه‌ها



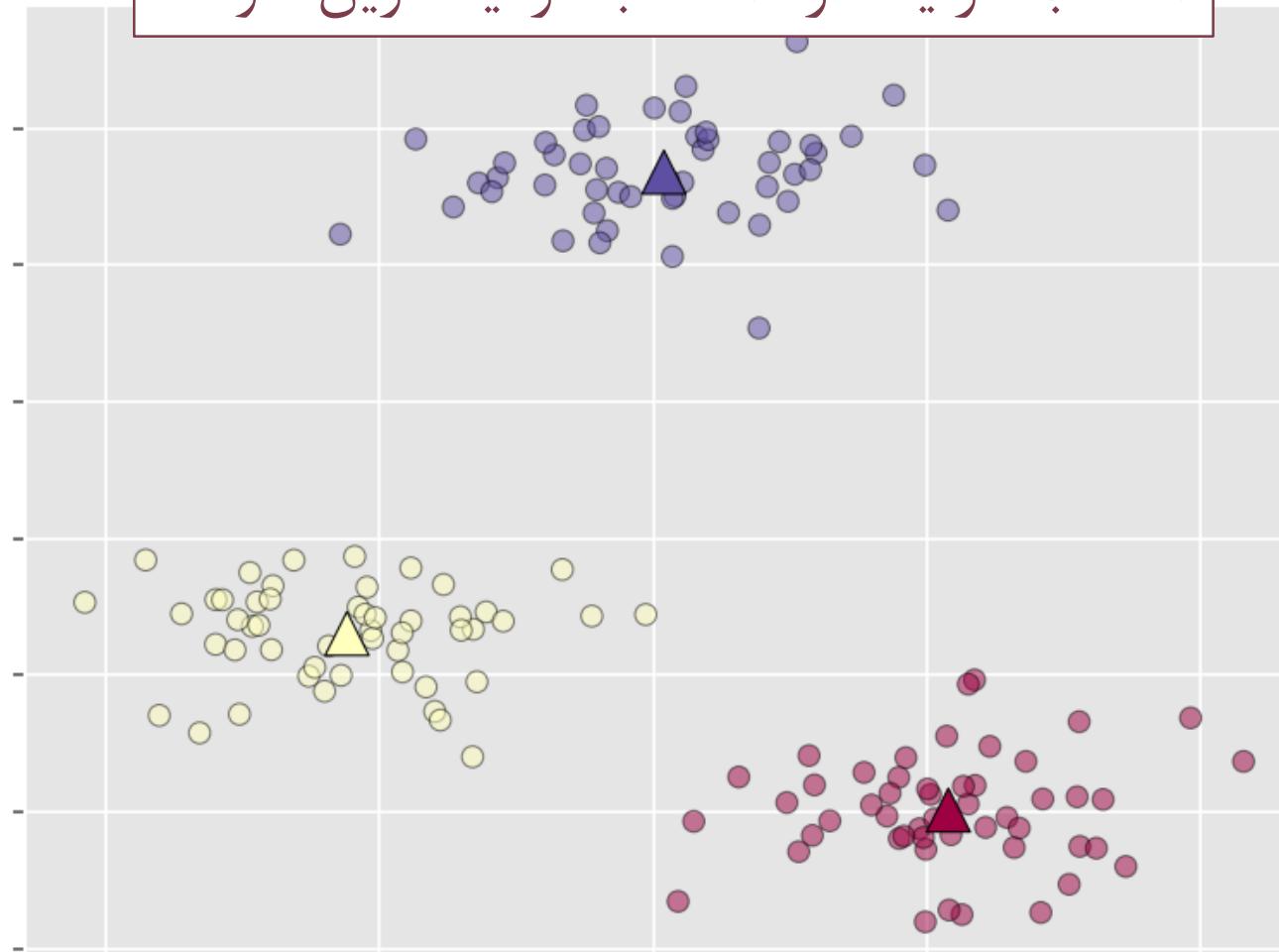
خوشه‌بندی: اجرای نمایشی

۱۹

انتساب هر یک از داده‌ها به نزدیک‌ترین خوشه

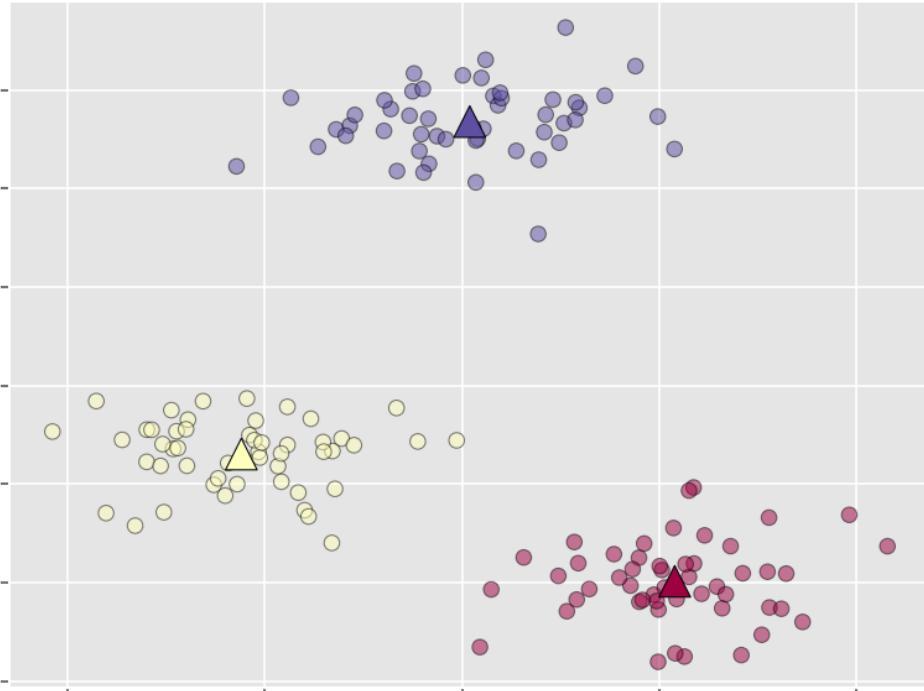
همگرایی:

خوشه هیچ یک از داده‌ها تغییر نکرد



الگوریتم K-means

۲۰



وروادی‌ها.

تعداد خوش‌های K :

مجموعه آموزشی:

$$\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(m)}\}$$

- توجه. در مجموعه آموزشی، هیچ برچسبی برای داده‌ها تعیین نشده است.
- توجه. در خوشبندی نیازی به افزودن ویژگی $x_0 = 1$ نیست.

الکوریتم K-means

۲۱

randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$

repeat

{

for $i = 1$ **to** m

انتساب داده‌ها به فوشهای

$$c^{(i)} = \arg \min_k \|x^{(i)} - \mu_k\|$$

for $k = 1$ **to** K

بروز رسانی مرکز فوشهای

μ_k = average of points assigned to cluster k

}

الکوئینٹس K-means

٢٢

```
centroids = np.random.random( (K, n) )

while True:

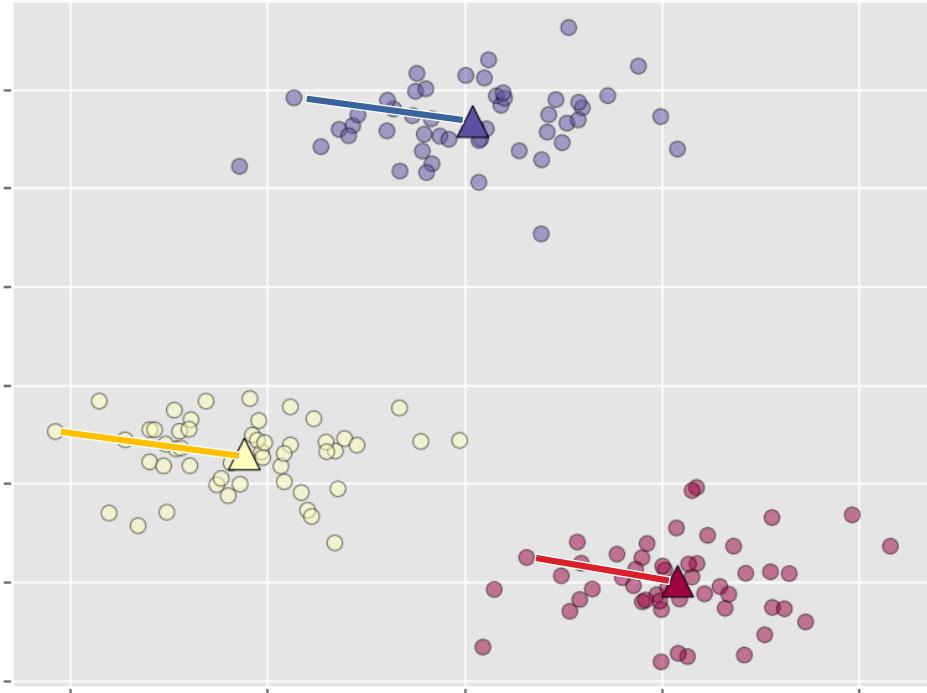
    for i in range(m):
        c[i] = np.argmin(np.linalg.norm(X[i] - centroids, axis=1))

    for k in range(K):
        centroids[k] = np.mean(X[c == k], axis=0)
```

خوشنندی: تابع هدف

تابع هدف

۲۴



نمادها.

μ_k : مرکز خوش \square

$x^{(i)}$: شماره خوش اختصاص یافته به داده \square

$x^{(i)}$: مرکز خوش اختصاص یافته به داده \square

تابع هدف.

$$J(c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_k) = \frac{1}{m} \sum_{i=1}^m \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

الگوریتم K-means

۲۵

randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$

repeat

{

for $i = 1$ to m

$$c^{(i)} = \arg \min_k \|x^{(i)} - \mu_k\|$$

کمینه‌سازی تابع هدف
نسبت به پارامترهای $c^{(i)}$

for $k = 1$ to K

μ_k = average of points assigned to cluster k

کمینه‌سازی تابع هدف
نسبت به پارامترهای μ

}

مقداردهی اولیه به مراکز خوشنده

الکوریتم K-means

٢٧

randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$

repeat

{

for $i = 1$ **to** m

$$c^{(i)} = \arg \min_k \|x^{(i)} - \mu_k\|$$

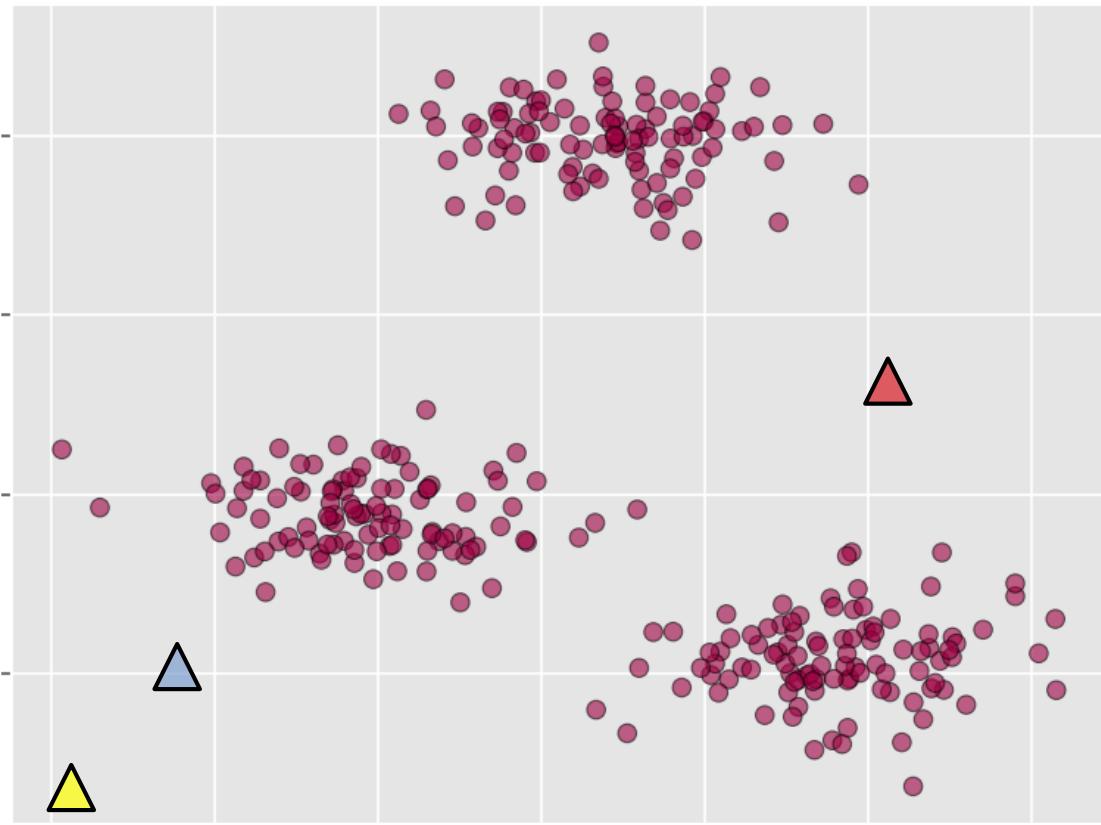
for $k = 1$ **to** K

μ_k = average of points assigned to cluster k

}

مدادهای اولیه به مرکز خوشها

۲۸

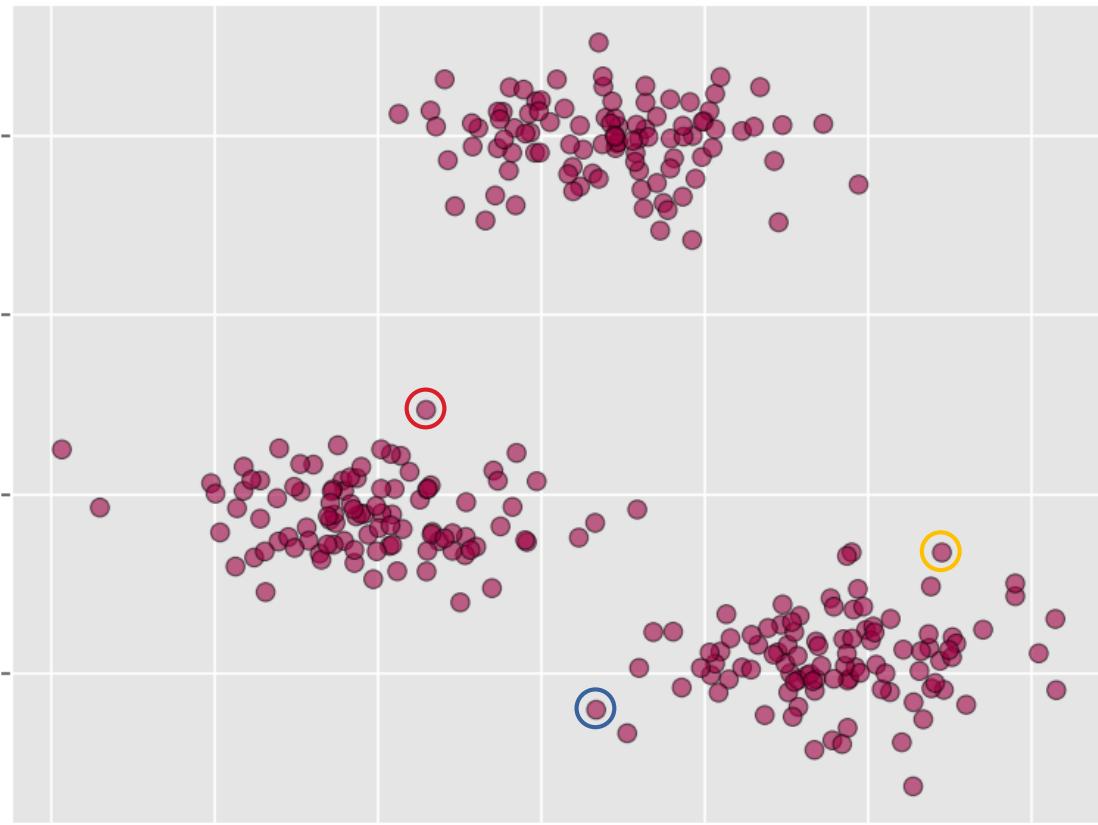


- مقداردهی اولیه. $[K \leq m]$
- انتخاب K نقطه به صورت تصادفی
- انتساب مراکز خوشها به K نقطه انتخاب شده

ممکن است یک مرکز به گونه‌ای انتخاب گردد که هیچ داده‌ای به آن تعلق نگیرد.

مدادهای اولیه به مرکز خوشها: (وش بهتر)

۲۹



□ مقداردهی اولیه. $[K \leq m]$

□ انتخاب K نمونه آموزشی به صورت تصادفی

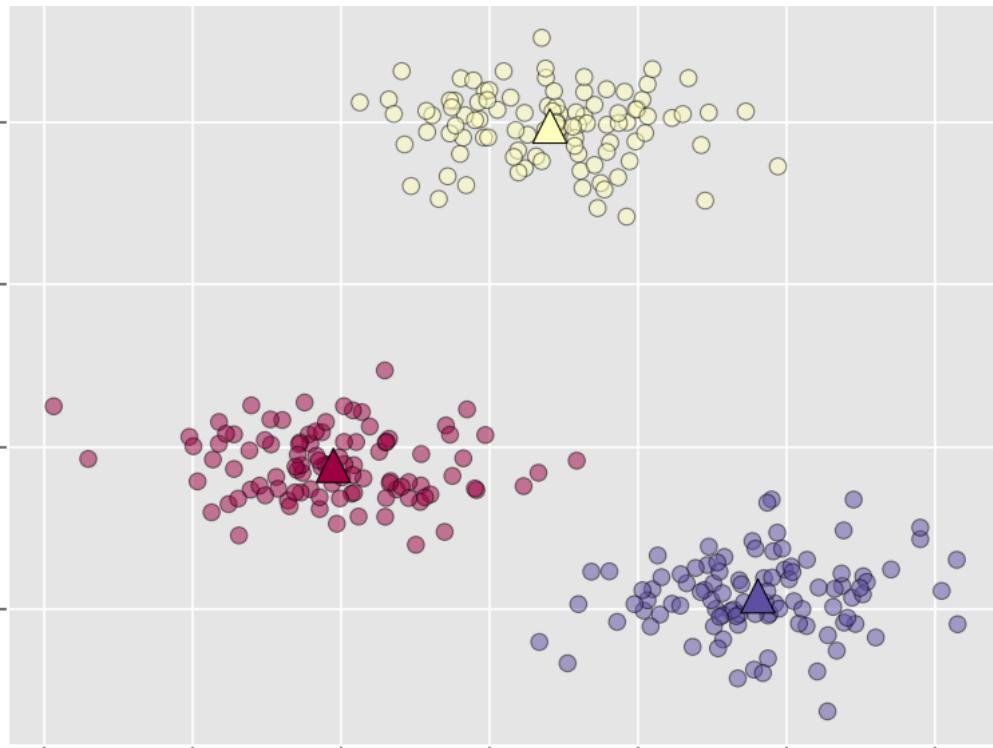
□ انتساب مراکز خوشها به K نمونه انتخاب شده

```
C = np.random.permutation(X) [:K]
```

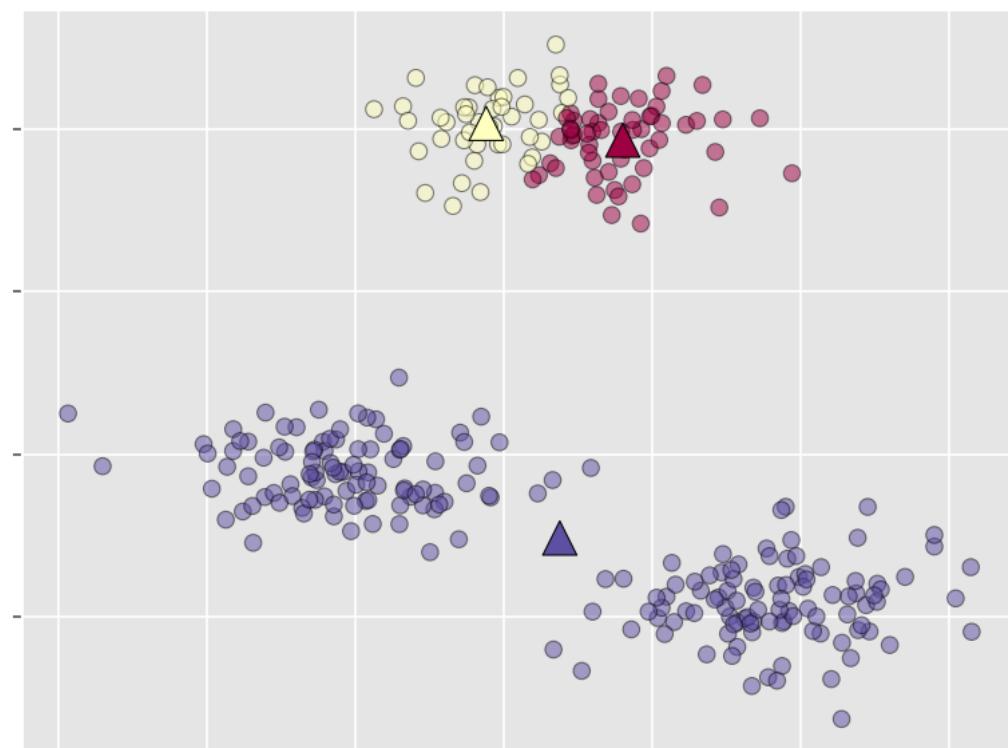
بهینه محلی و بهینه سراسری

۳۰

بهینه سراسری



بهینه محلی



اجتناب از بهینه‌های محلی

۳۱

for $t = 1$ **to** MAX

{

randomly initialize cluster centroids $\mu_1, \mu_2, \dots, \mu_k$

run K-means to get $c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_k$

compute cost function $J(c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_k)$

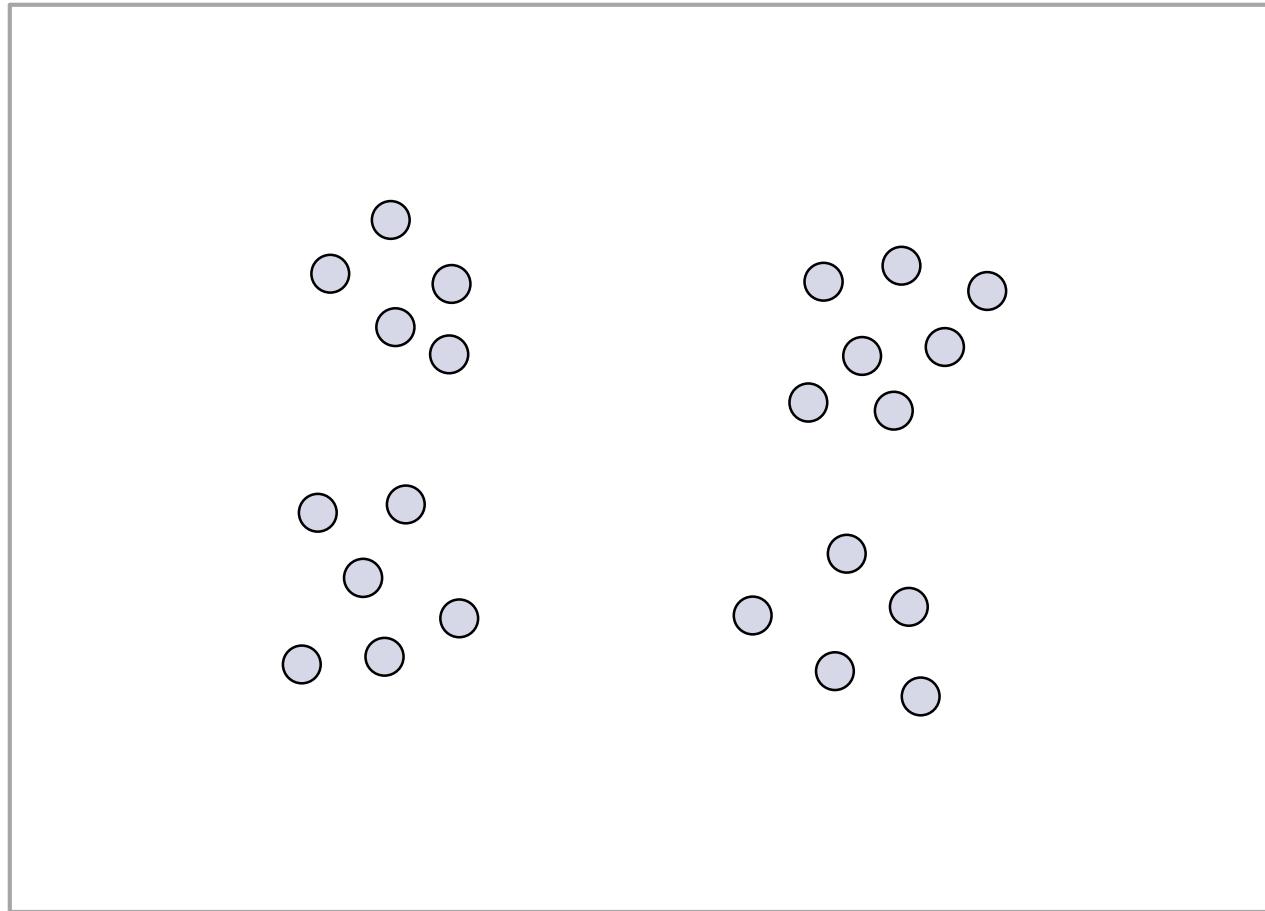
}

pick clustering with minimum cost

تعین تعداد خوشنما

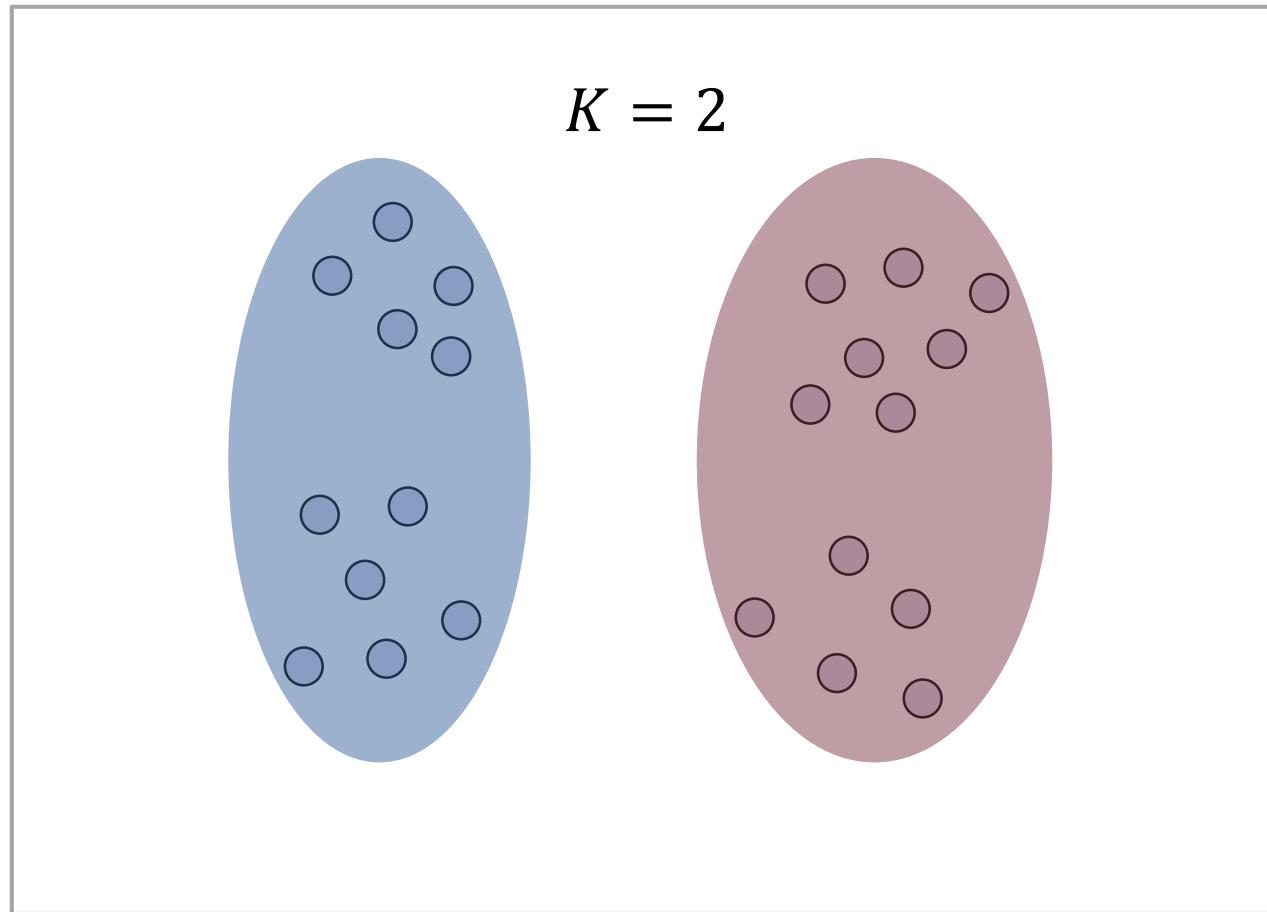
مدقار مناسب برای K گداخ است؟

۳۳



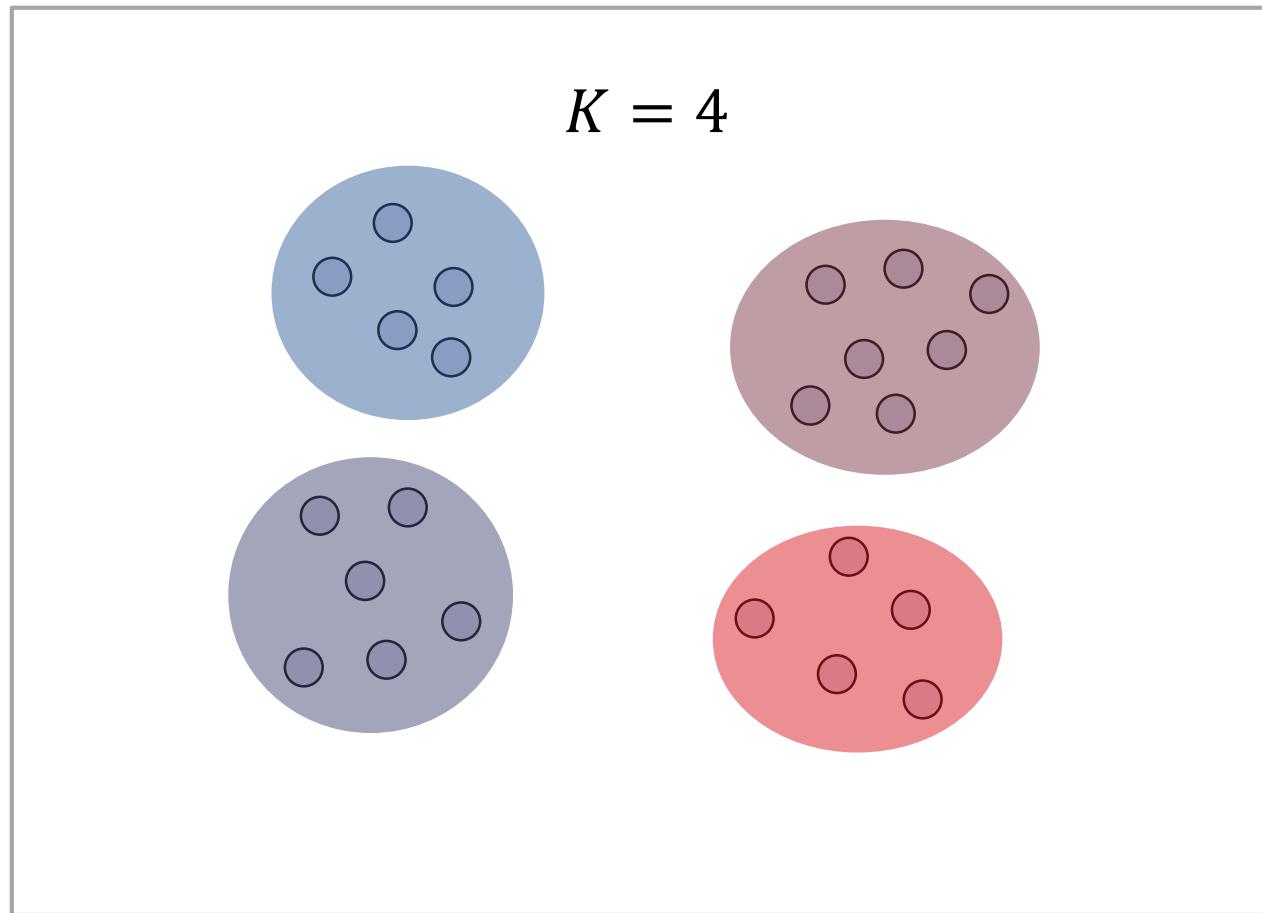
مدادار مناسب برای K گدام است؟

۳۴



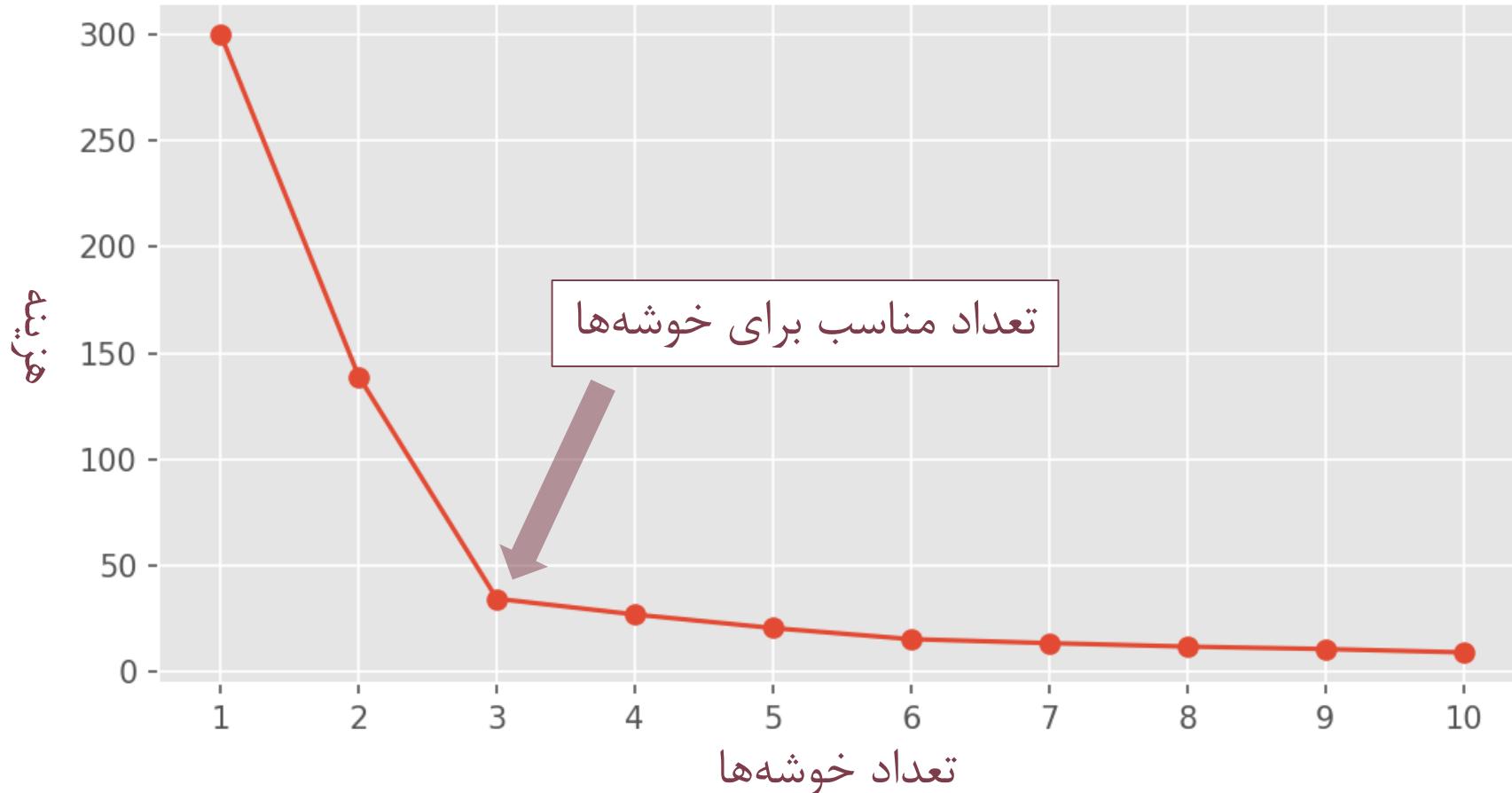
مدقار مناسب برای K گذاخت است؟

۳۵



تعیین تعداد مناسب خوشه‌ها: (وش آرنه)

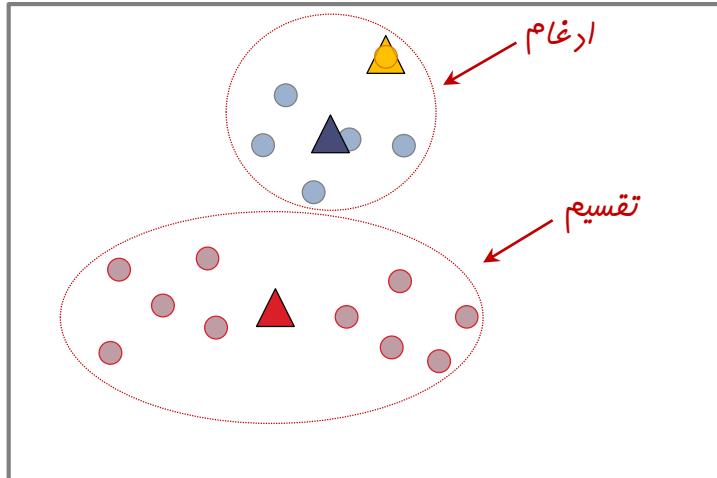
۳۶



برهیود خوشنندی

بهبود خوشه‌بندی با پس‌پردازش خوشه‌ها

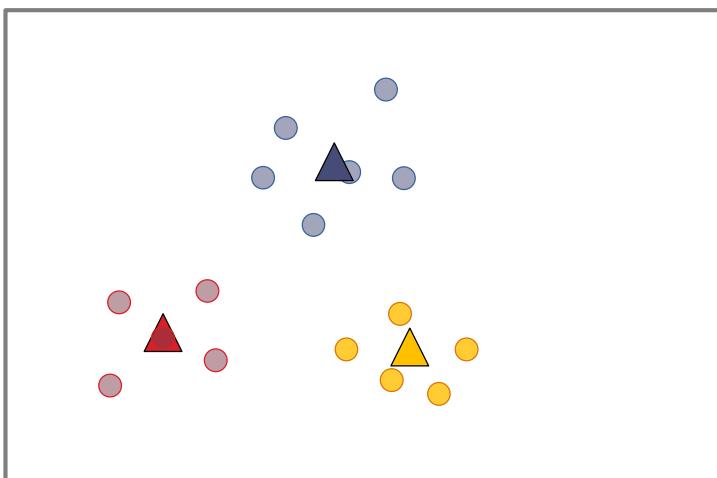
۳۸



□ تقسیم یک خوشه با بیشترین خطا به دو خوشه

با اجرای K-means بر روی داده‌های این خوشه با مقدار $K = 2$

□ تقسیم.



□ ادغام نزدیک‌ترین دو خوشه

□ ادغام دو خوشه با حداقل افزایش در مجموع خطا

□ ادغام.

الگوریتم K-means دو بخشی‌ساز

۳۹

- الگوریتم دو بخشی‌ساز.
- با یک خوشه شامل تمامی داده‌ها شروع کن.
- هر بار یک خوشه را انتخاب کن:
 - خوشه انتخاب شده را به وسیله الگوریتم K-means به دو خوشه تقسیم کن.
 - مجموع خطای خوشبندی را محاسبه کن.
 - خوشبندی با کمترین خطا را انتخاب کن.
- عمل بالا را تا زمان رسیدن به تعداد خوشه‌های مورد نظر تکرار کن.

الگوریتم K-means دو بخشی‌ساز

۴۰

Start with all the points in one cluster

while the number of clusters is less than K

measure the total error

for every cluster

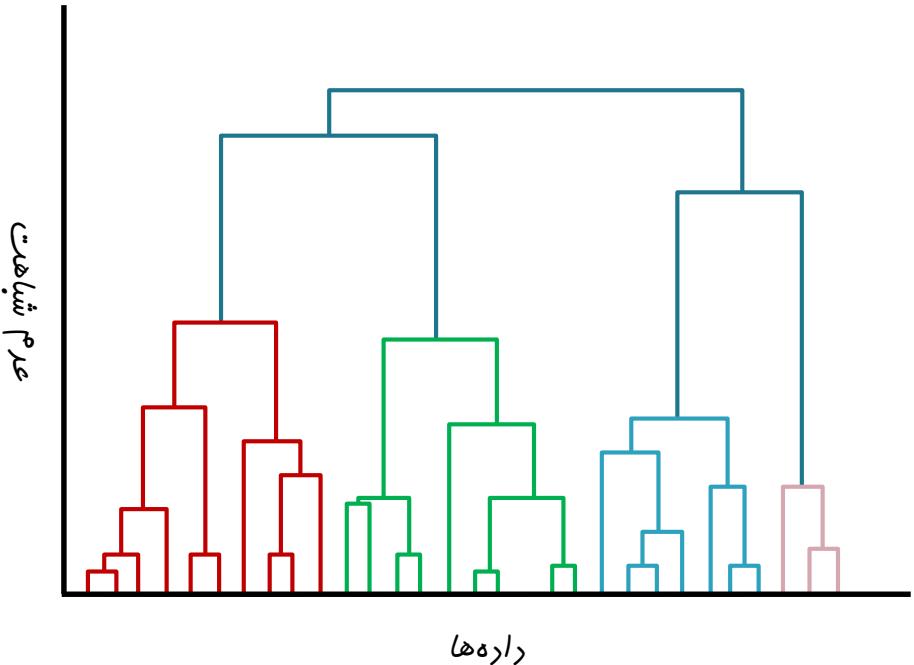
perform K-means clustering with $k = 2$ on the given cluster

measure the total error after splitting

choose the cluster split that gives the lowest error

خوشه‌بندی سلسله‌مراتبی

۴۱



□ ایجاد یک **درخت نگاره** شامل یک طیف گسترده از خوشه‌بندی‌ها.

□ خوشه‌بندی سلسله‌مراتبی.

□ ابتدا داده‌های بسیار شبیه را ادغام کن.

□ به تدریخ با ادغام خوشه‌های کوچک‌تر، خوشه‌های بزرگ‌تری ایجاد کن.

□ الگوریتم.

□ در ابتدا هر داده بیانگر یک خوشه است.

□ مراحل زیر را تکرار کن:

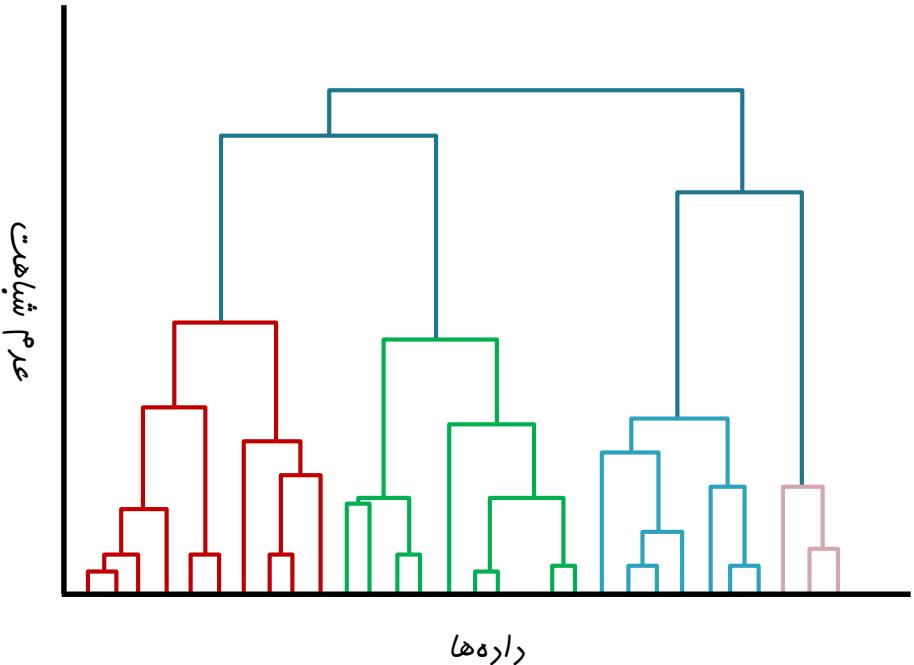
■ هر بار **نزدیک‌ترین** دو خوشه را انتخاب کن.

■ آن دو خوشه را در یک خوشه جدید ادغام کن.

■ توقف: زمانی که تنها یک خوشه باقی مانده باشد.

خوشه‌بندی سلسله‌مراتبی

۴۲



□ ایجاد یک **درخت نگاره** شامل یک طیف گسترده از خوشه‌بندی‌ها.

□ خوشه‌بندی سلسله‌مراتبی.

□ ابتدا داده‌های بسیار شبیه را ادغام کن.

□ به تدریخ با ادغام خوشه‌های کوچک‌تر، خوشه‌های بزرگ‌تری ایجاد کن.

□ الگوریتم.

□ در ابتدا هر داده بیانگر یک خوشه است.

□ مراحل زیر را تکرار کن:

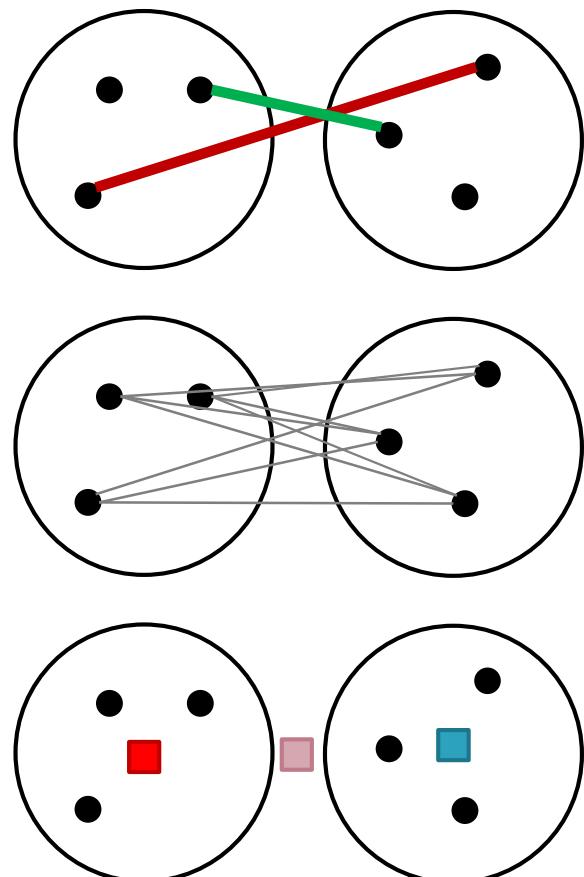
■ هر بار **نزدیک‌ترین** دو خوشه را انتخاب کن.

■ آن دو خوشه را در یک خوشه جدید ادغام کن.

■ توقف: زمانی که تنها یک خوشه باقی مانده باشد.

خوشه‌بندی سلسله‌مراتبی

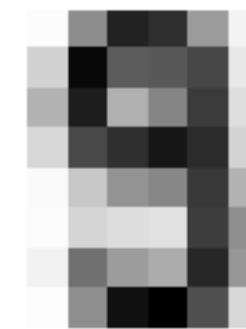
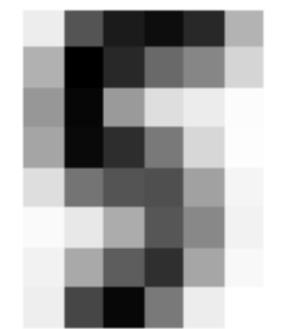
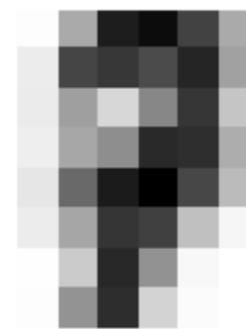
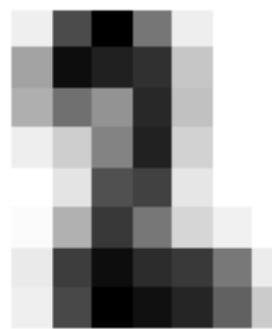
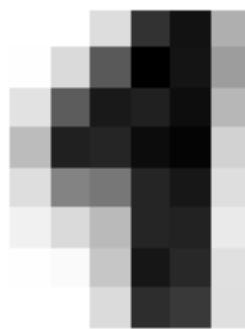
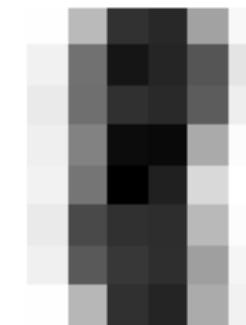
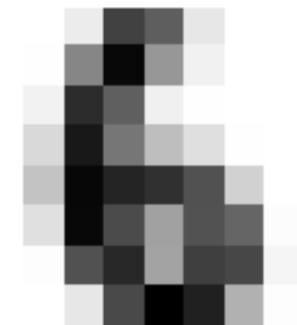
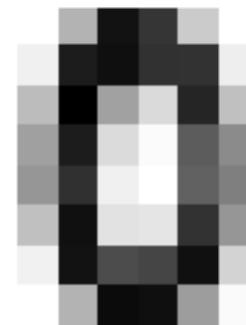
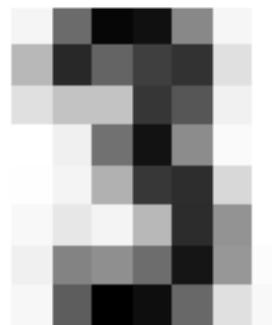
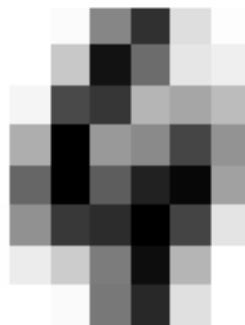
۴۳



- چگونه می‌توان نزدیک‌ترین دو خوشه را تعریف نمود?
- معیارهای تعیین شباهت خوشه‌ها.
- نزدیک‌ترین زوج (خوشه‌بندی تک-پیوندی)
- دورترین زوج (خوشه‌بندی تمام-پیوندی)
- میانگین فاصله همه زوج‌ها
- روش «وارد» (کمترین پراکندگی، مانند k-means)
- معیارهای مختلف باعث ایجاد خوشه‌بندی‌های متفاوتی می‌شوند.

خوشه‌بندی ارقام

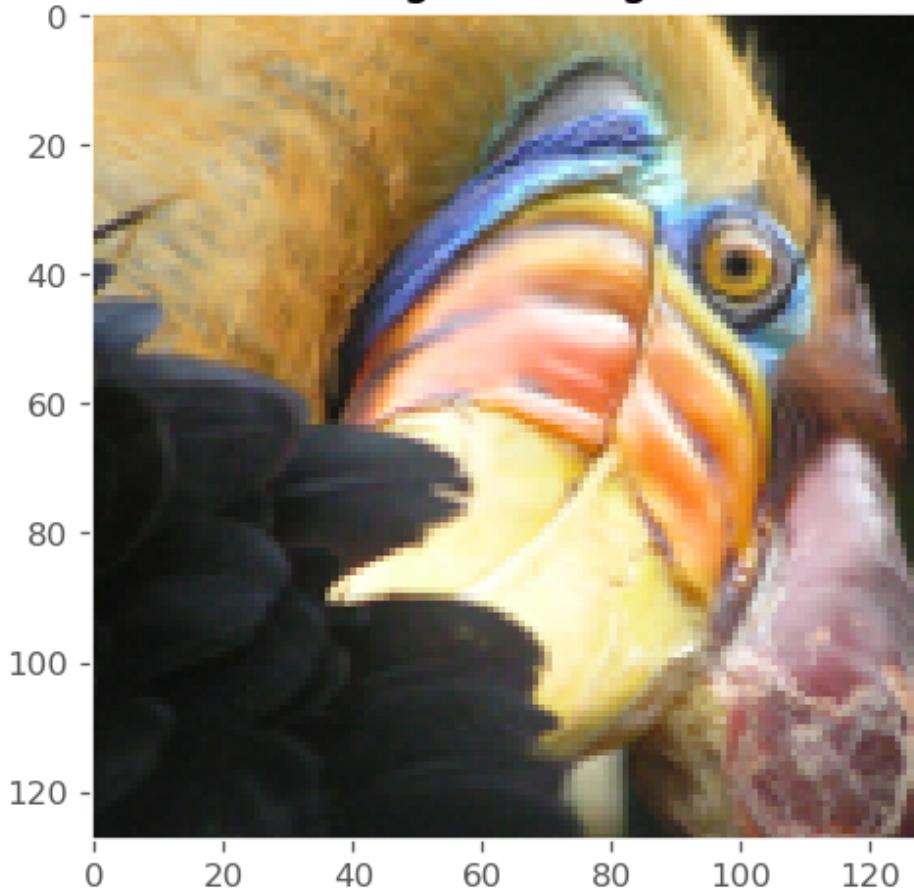
۴۴



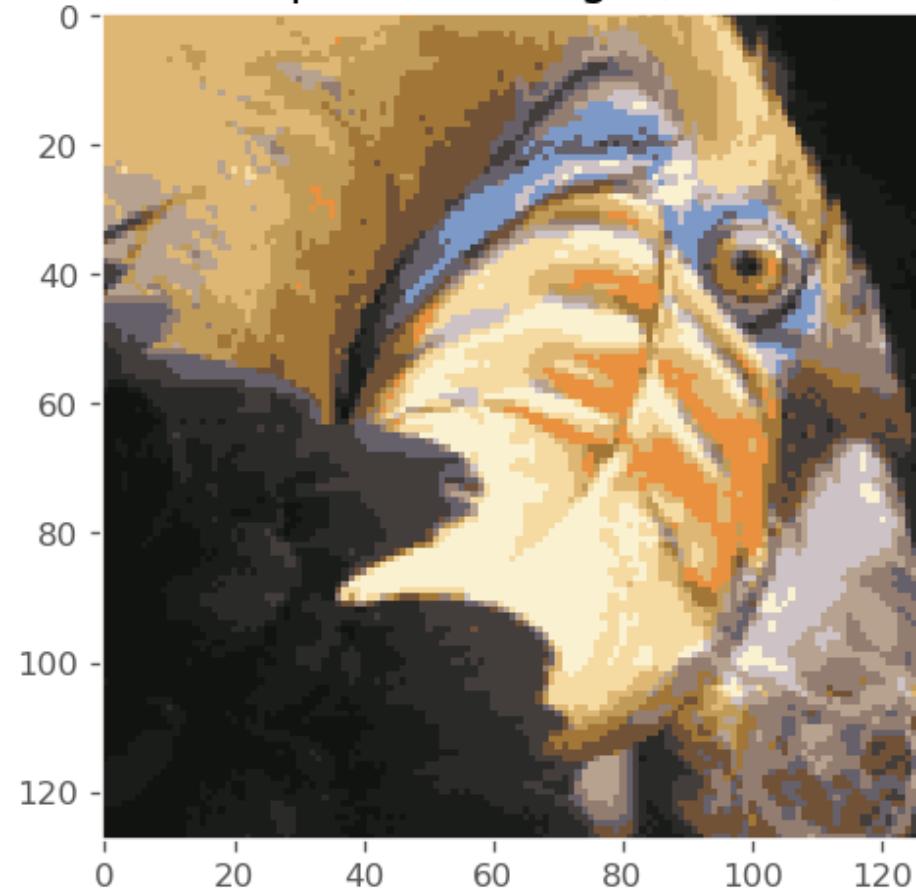
فُشرده‌سازی تصویر

۴۵

Original Image

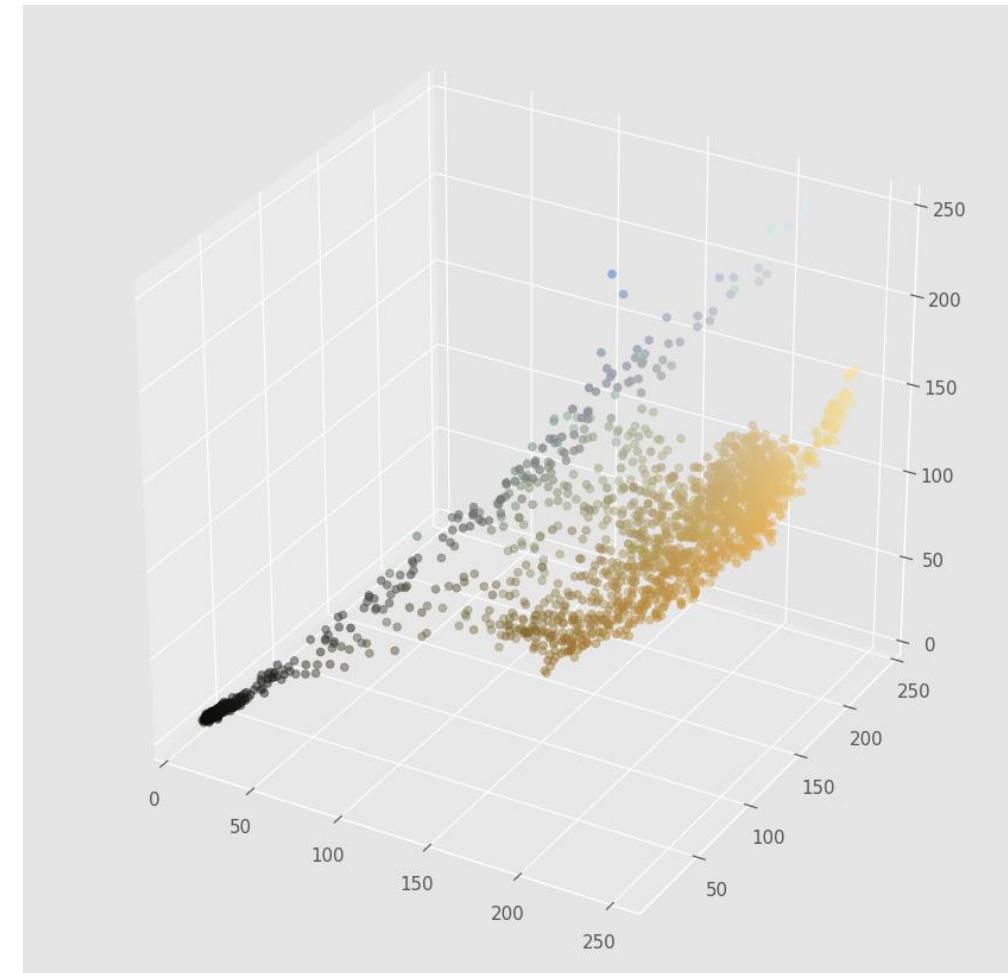
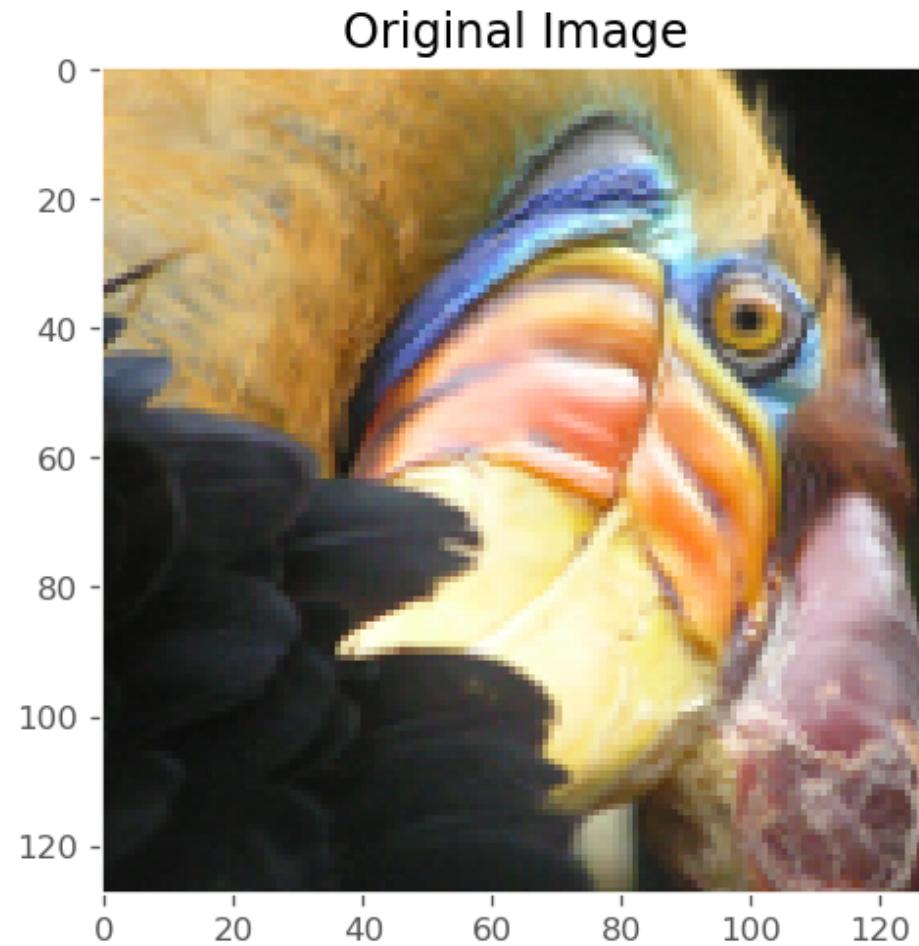


Compressed Image ($K = 16$)



فُلْسِرِدَه سازی تَصویر

۴۶



خلاصه

۴۷

- یادگیری بدون نظارت. یافتن ساختار در داده‌ها
- خوشه‌بندی. گروه‌بندی داده‌های مشابه
 - ▣ الگوریتم خوشه‌بندی K-means
 - پیاده‌سازی آسان
 - برای مجموعه داده‌های بسیار بزرگ کند
 - امکان گیر کردن در بهینه محلی
 - ▣ پس‌پردازش خوشه‌ها: تقسیم و ادغام خوشه‌ها
- الگوریتم K-means دو بخشی‌ساز
 - خوشه‌بندی بهتر نسبت به الگوریتم K-means
- الگوریتم‌های خوشه‌بندی سلسله‌مراتبی

کاهش ابعاد

سید ناصر رضوی www.snrazavi.ir

۱۳۹۷

کاهش ابعاد

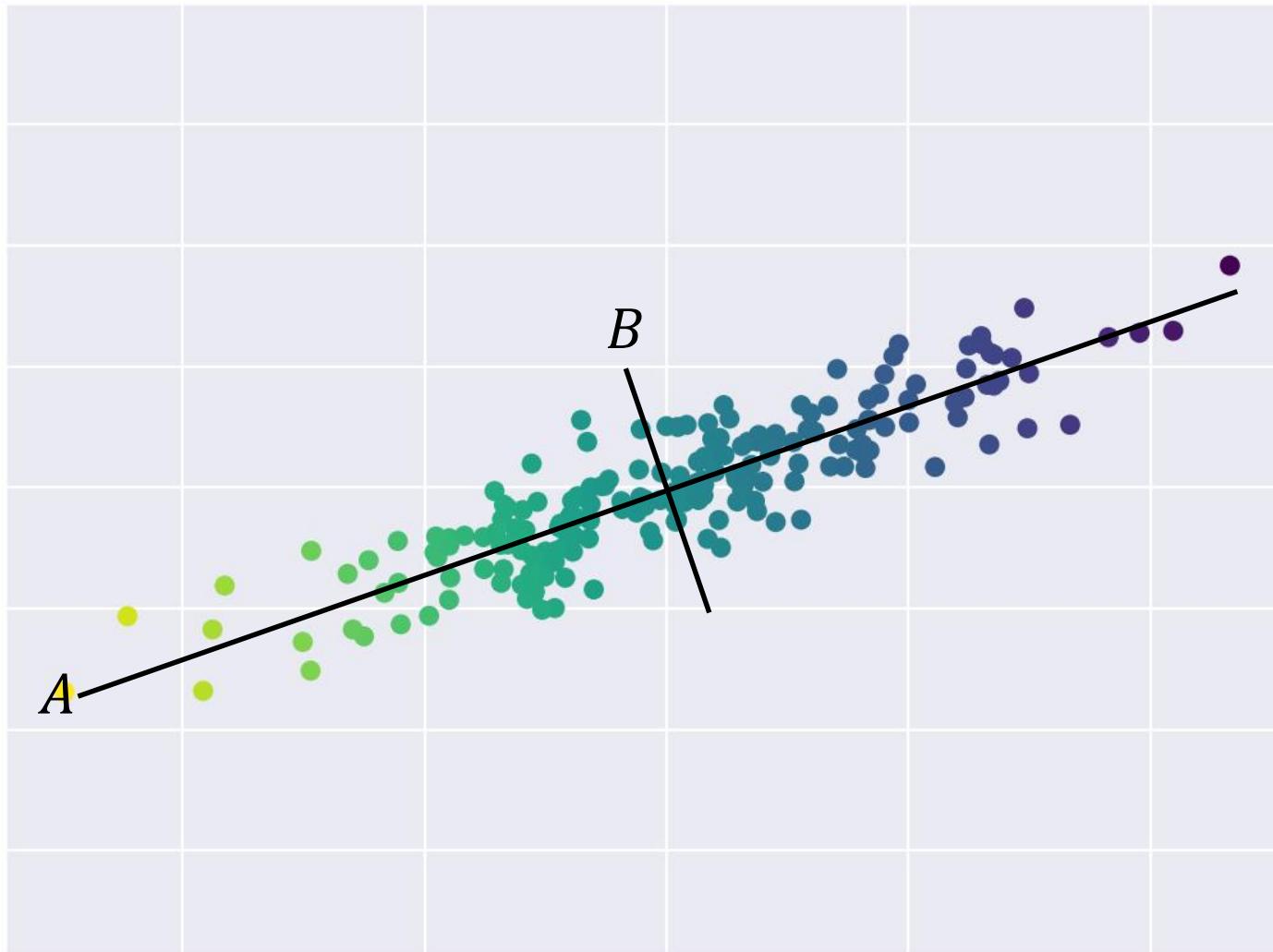
۲

- روش‌ها.
- تحلیل مؤلفه‌های اصلی
- تحلیل فاکتورها
- تحلیل مؤلفه‌های مستقل
- انگیزه.
- نمایش داده‌ها
- فشرده‌سازی داده‌ها
- کاهش مصرف حافظه
- استفاده آسان‌تر از مجموعه داده‌ها
- کاهش هزینه‌های محاسباتی بسیاری از الگوریتم‌ها
- حذف نویز و افزایش دقیقیت الگوریتم یادگیری
- ساده‌تر کردن درک نتایج

تملیل مؤلفه‌های اصلی (PCA)

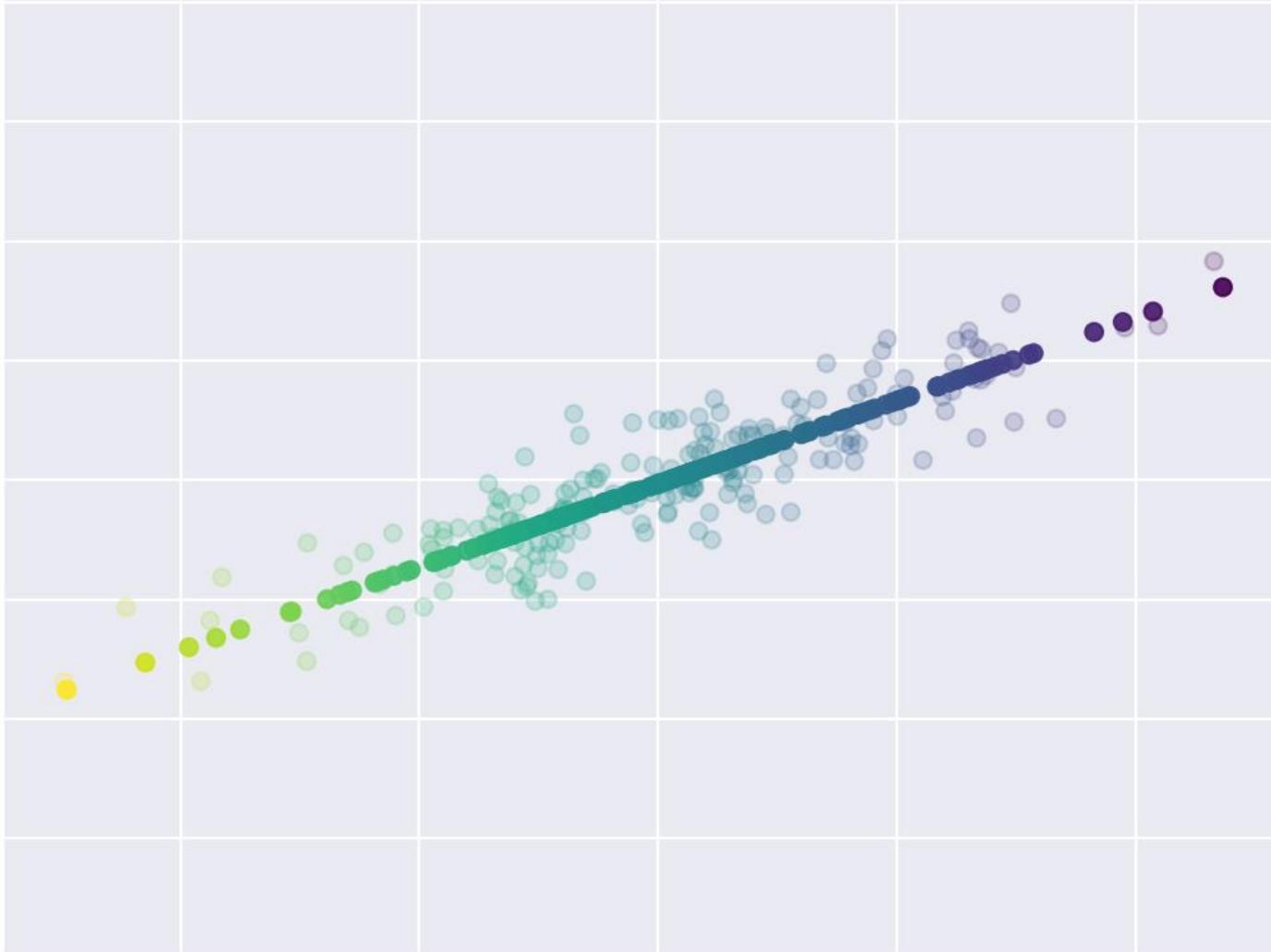
تمثیل مؤلفه‌های اصلی

۴



تمثیل مؤلفه‌های اصلی

۵

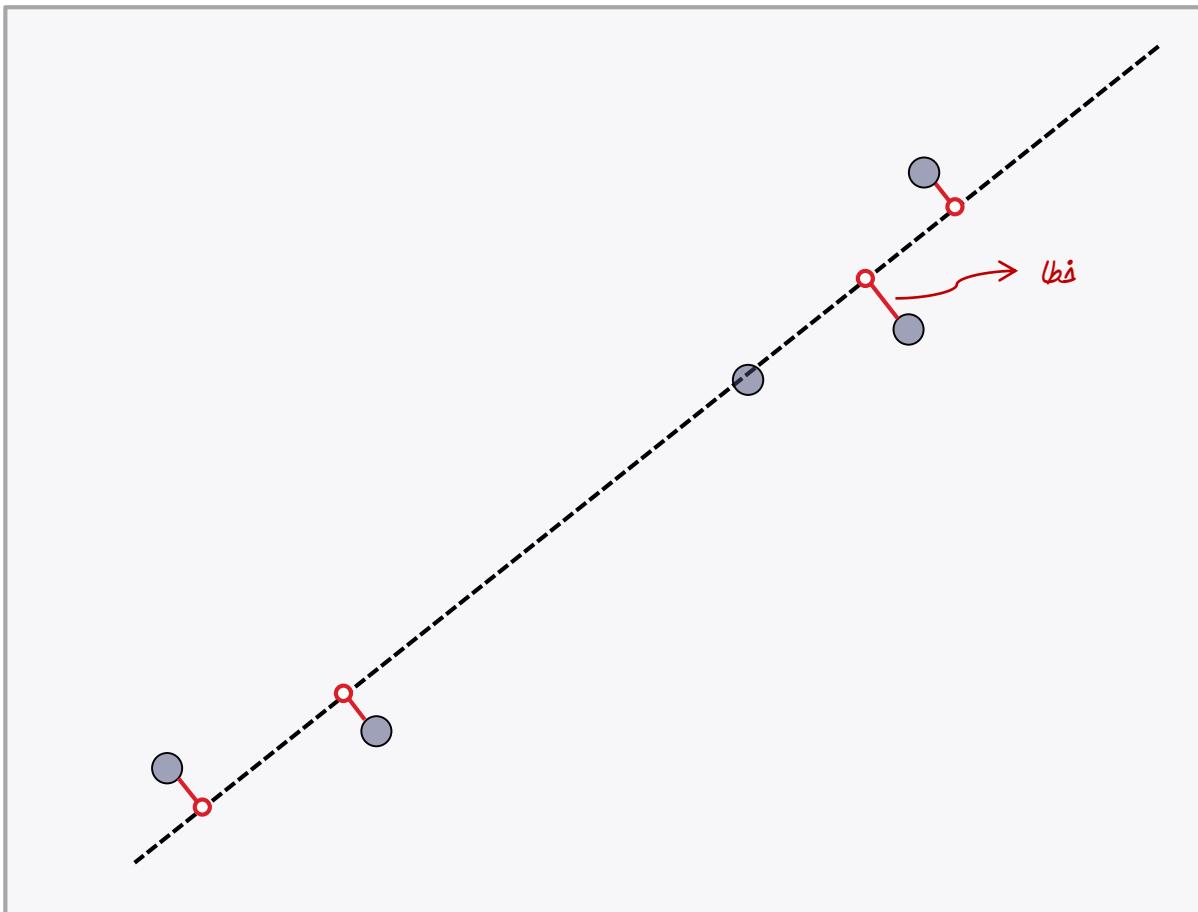


تملیل مؤلفه‌های اصلی: بیان مسئله

بیان مسئله

۷

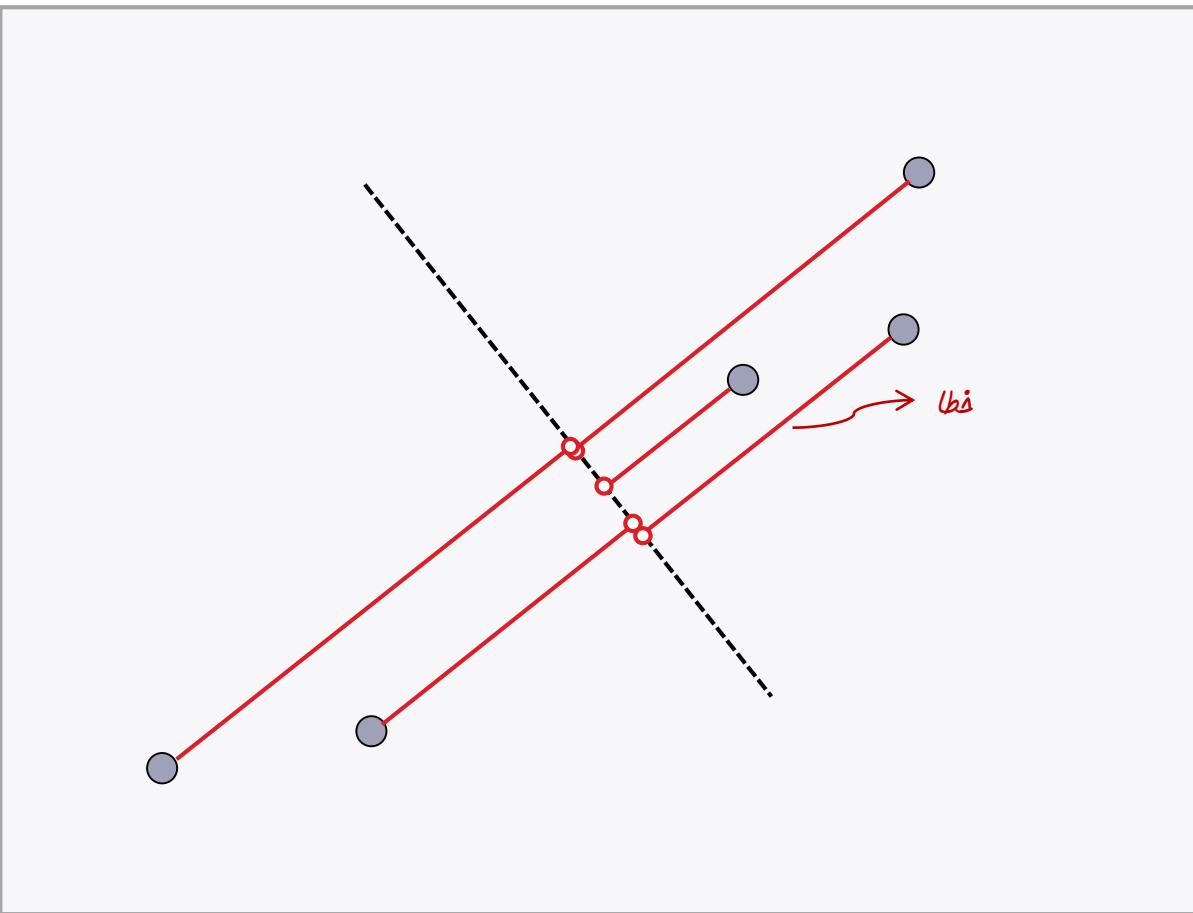
□ هدف. کمینه‌سازی مجموع مربعات خطای تابش



بیان مسئله

۸

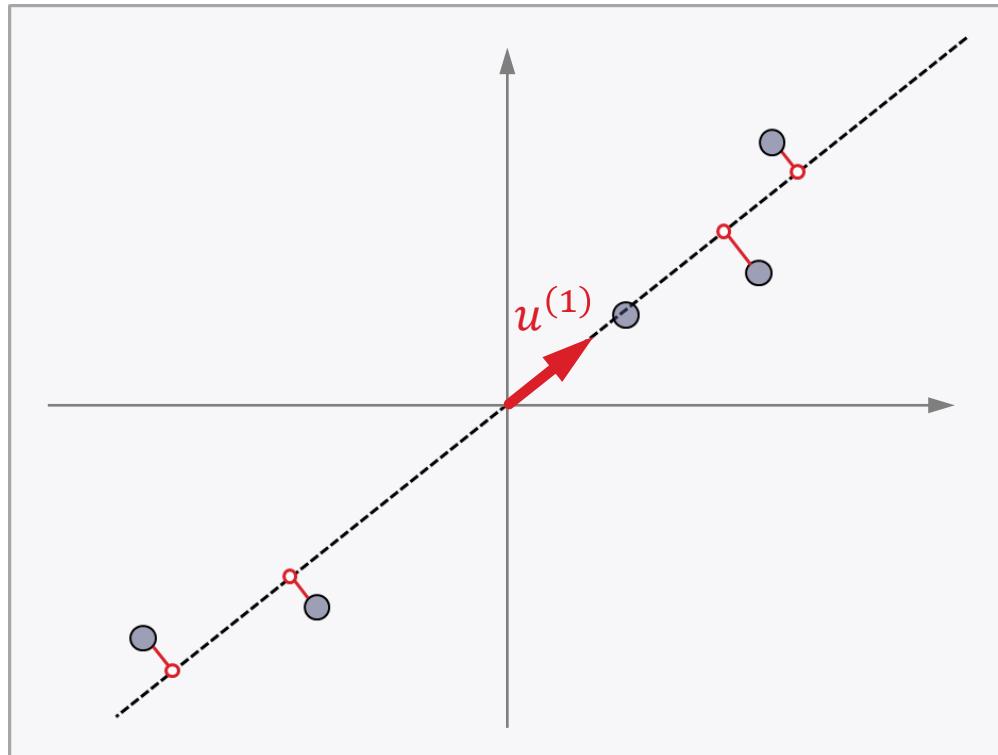
□ هدف. کمینه‌سازی مجموع مربعات خطای تابش



بیان مسئله

۹

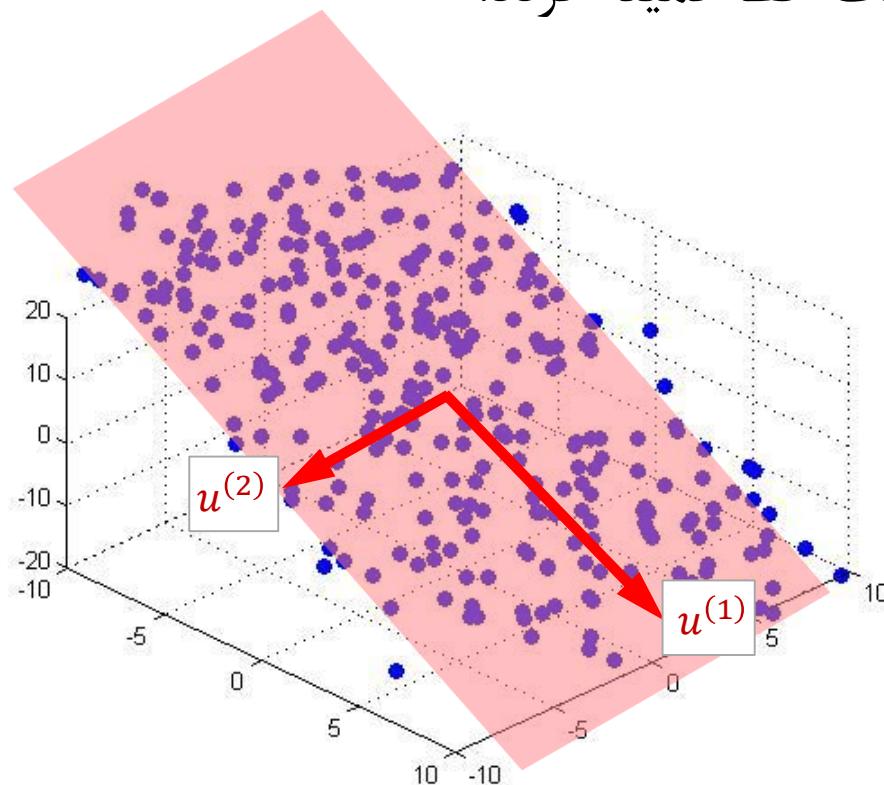
□ کاهش ابعاد از ۲ به ۱. یافتن یک جهت مانند $u^{(1)} \in \mathbb{R}^2$ به طوری که با تصویر کردن نقاط در آن جهت، مجموع مربعات خطای کمینه گردد.



بیان مسئله

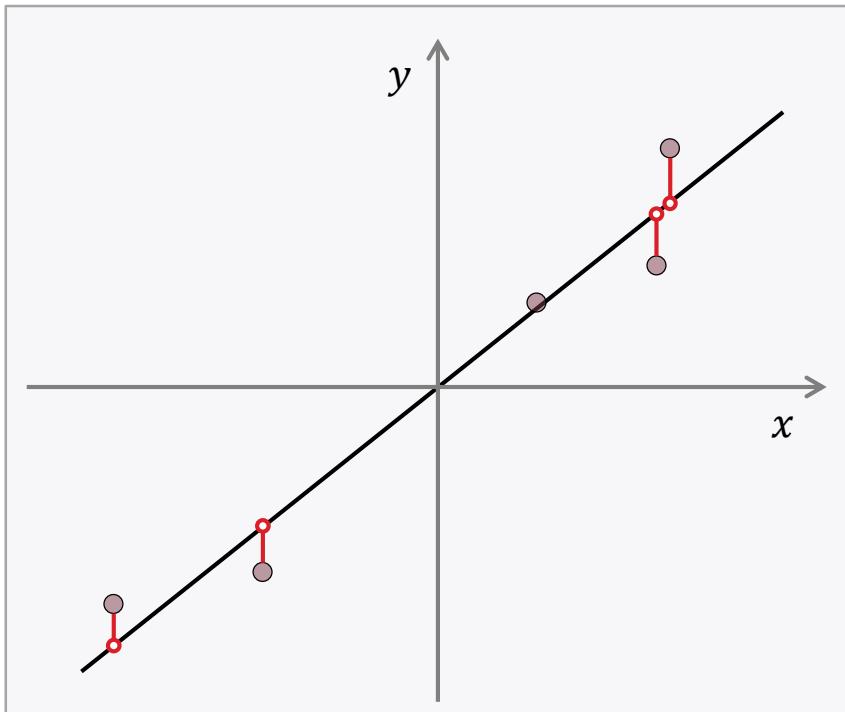
۱۰

- کاهش ابعاد از n به k . یافتن k بردار متعامد مانند $u^{(1)}, u^{(2)}, \dots, u^{(k)} \in \mathbb{R}^n$ به طوری که با تصویر کردن نقاط در آن جهت‌ها، مجموع مربعات خطای کمینه گردد.

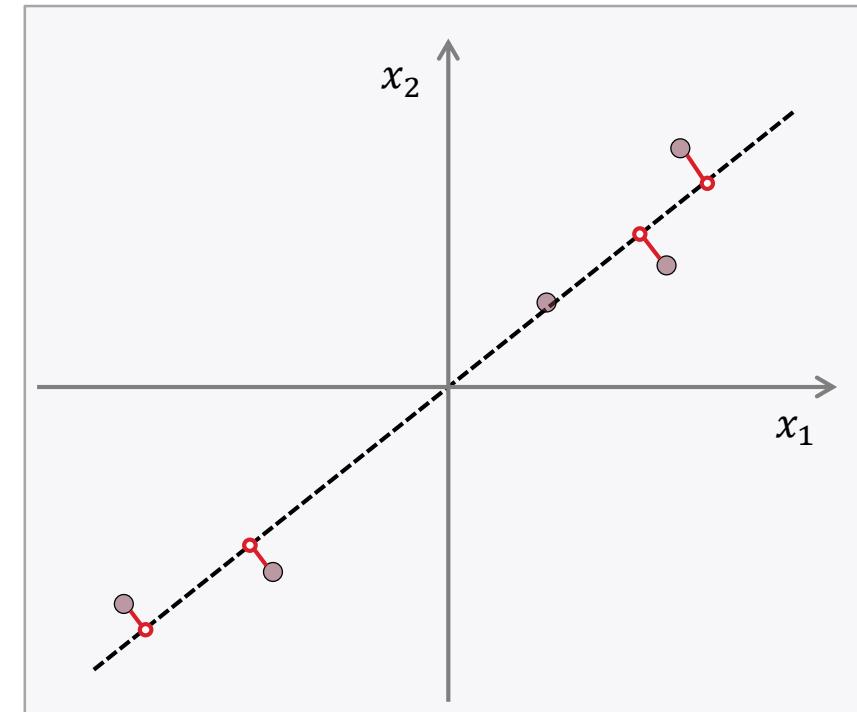


همان رگرسیون است؟ PCA

۱۱



رگرسیون خطی



تحلیل مؤلفه‌های اصلی

الگوریتم PCA

الگوريتم PCA: پيشپردازش

۱۳

□ مجموعه آموزشی.

$$\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(m)}\}, \quad x^{(i)} \in \mathbb{R}^n$$

□ پيشپردازش.

(۱) حذف ميانگين.

$$\mu_j = \frac{1}{m} \sum_{i=1}^m x_j^{(i)}, \quad x_j^{(i)} = x_j^{(i)} - \mu_j$$

(۲) مقیاسبندی. [در صورت نياز]

$$x_j^{(i)} = \frac{x_j^{(i)} - \mu_j}{s_j}$$

انحراف معيار در ويزگي j

الگوریتم PCA: کاهش ابعاد

۱۴

$$\Sigma = \frac{1}{m} X^T X = \frac{1}{m} \sum_{i=1}^n x^{(i)} (x^{(i)})^T$$

$$[U, S, V] = svd(\Sigma)$$

$$U = \begin{bmatrix} | & | & \cdots & | \\ u^{(1)} & u^{(2)} & \cdots & u^{(n)} \\ | & | & \cdots & | \end{bmatrix}_{n \times n} \quad \Rightarrow$$

$$U_{reduced} = \begin{bmatrix} | & | & \cdots & | \\ u^{(1)} & u^{(2)} & \cdots & u^{(k)} \\ | & | & \cdots & | \end{bmatrix}_{n \times k}$$

□ کاهش ابعاد داده‌ها از n به k .

□ محاسبه «ماتریس کوواریانس»:

□ محاسبه «بردارهای ویژه» ماتریس کوواریانس:

□ انتخاب k بردار اول از ماتریس U :

الگوریتم PCA

۱۵

محاسبه داده‌های جدید با ابعاد $.k$ □

$$z_{k \times 1}^{(i)} = \begin{bmatrix} | & | & \cdots & | \\ u^{(1)} & u^{(2)} & \dots & u^{(k)} \\ | & | & \cdots & | \end{bmatrix}^T \times x_{n \times 1}^{(i)}$$

$$= \begin{bmatrix} - & u^{(1)} & - \\ - & u^{(2)} & - \\ \vdots & \vdots & \vdots \\ - & u^{(k)} & - \end{bmatrix}_{k \times n} \times x_{n \times 1}^{(i)}$$

پیاده سازی در پایتون

۱۶

□ الگوریتم PCA.

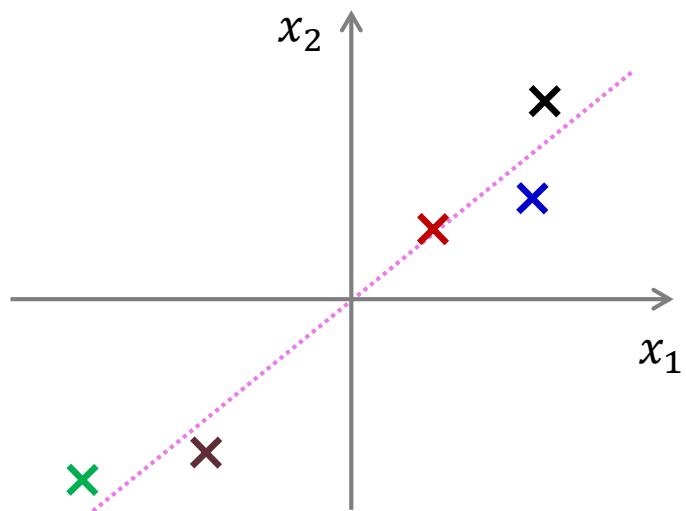
□ پس از حذف میانگین و در صورت نیاز مقیاس بندی.

```
def PCA(X, k):  
  
    m = X.shape[0]  
  
    Sigma = (X.T @ X) / m  
    ← محاسبه ماتریس کوواریانس  
  
    U, S, V = svd(Sigma)  
    ← محاسبه تجزیه مقادیر منفرد  
  
    U_reduced = U[:, :k]  
    ← انتخاب  $k$  مؤلفه اول  
  
    Z = X @ U_reduced  
    ← محاسبه داده های جدید با  $k$  بعد  
  
    return Z
```

مثال: کاهش ابعاد

۱۷

$$Z = X \times U_{reduced}$$



داده‌های اولیه (دو بعدی)

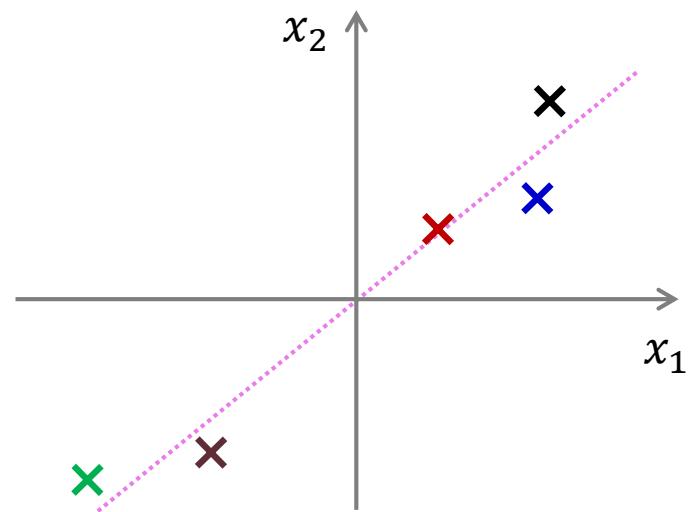


داده‌های جدید (یک بعدی)

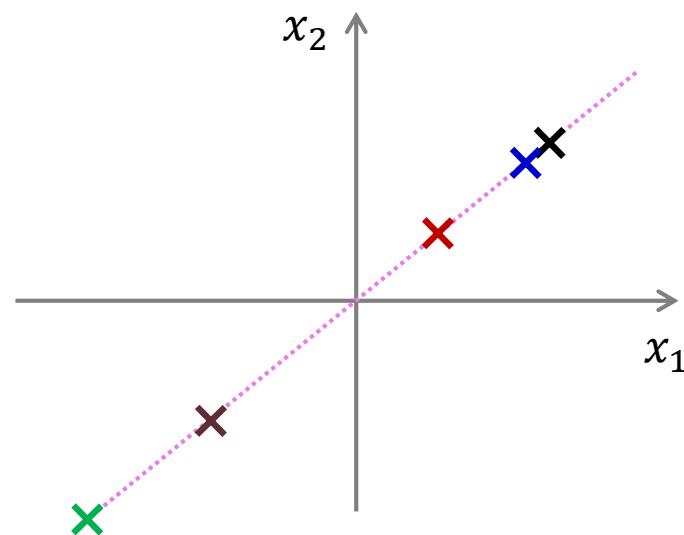
مثال: کاهش ابعاد

۱۸

$$X_{recovered} = Z * U_{reduced}^T + means$$



داده‌های اولیه (دو بعدی)



داده‌های بازسازی شده

الگوریتم PCA: مذف میانگین

۱۹

```
x = np.array([[1, 1, 1, 0, 0],  
              [2, 2, 2, 0, 0],  
              [1, 1, 1, 0, 0],  
              [5, 5, 5, 0, 0],  
              [1, 1, 0, 2, 2],  
              [0, 0, 0, 3, 3],  
              [0, 0, 0, 1, 1]])
```

```
mu = X.mean(axis=0)  
X_norm = X - mu
```

دراجه های اولیه

1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
1	1	0	2	2
0	0	0	3	3
0	0	0	1	1

الگوریتم PCA: مماسنی بردارهای ویژه

۲۰

```
m = X.shape[0]
Sigma = (X_norm.T @ X_norm) / m
U, S, V = np.linalg.svd(Sigma)
print(S)
```

$$S = \begin{bmatrix} s_{11} & 0 & \cdots & 0 \\ 0 & s_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_{nn} \end{bmatrix}$$

[8.72e+00 1.58e+00 6.69e-02 4.79e-16 1.35e-47]

الگوريتم PCA: کاهش ابعاد

۲۱

```
U_red = U[:, :3]
```

```
X_proj = X_norm * U_red
```

```
[[ 0.17   1.37   -0.01]
 [-1.44   0.74   -0.02]
 [ 0.17   1.37   -0.01]
 [-6.28  -1.17   -0.06]
 [ 1.76  -1.1     0.57]
 [ 3.33  -1.92  -0.35]
 [ 2.3    0.7    -0.11]]
```

الگوریتم PCA: بازسازی داده های اولیه

۲۲

```
x_approx = x_proj @ U_red + mu
```

```
[[ 1.00   1.00   1.00   0.00   0.00]
  2.00   2.00   2.00   0.00   0.00]
  1.00   1.00   1.00   0.00   0.00]
  5.00   5.00   5.00  -0.00  -0.00]
  1.00   1.00   0.00   2.00   2.00]
  0.00  -0.00   0.00   3.00   3.00]
 -0.00  -0.00  -0.00   1.00   1.00]]
```

داده های اولیه

1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
1	1	0	2	2
0	0	0	3	3
0	0	0	1	1

انتساب تعداد مؤلفهای اصلی

انتفاب تعداد مؤلفه‌های اصلی

۲۴

□ میانگین مجموع مربعات خطای تابش.

$$\frac{\frac{1}{m} \sum_{i=1}^m \|x^{(i)} - x_{approx}^{(i)}\|^2}{\frac{1}{m} \sum_{i=1}^m \|x^{(i)}\|^2} \leq 0.01$$

حفظ ۹۹ درصد واریانس

$$\frac{\frac{1}{m} \sum_{i=1}^m \|x^{(i)} - x_{approx}^{(i)}\|^2}{\frac{1}{m} \sum_{i=1}^m \|x^{(i)}\|^2} \leq 0.05$$

حفظ ۹۵ درصد واریانس

$$\frac{\frac{1}{m} \sum_{i=1}^m \|x^{(i)} - x_{approx}^{(i)}\|^2}{\frac{1}{m} \sum_{i=1}^m \|x^{(i)}\|^2} \leq 0.10$$

حفظ ۹۰ درصد واریانس

انتفاپ تعداد مؤلفه‌های اصلی

۲۵

$k = 0$

یک الگوریتم غیرکارا

repeat

{

$k = k + 1$

try $PCA(X)$ with k components

compute $U_{reduced}, Z^{(1)}, Z^{(2)}, \dots, Z^{(m)}, x_{approx}^{(1)}, x_{approx}^{(2)}, \dots, x_{approx}^{(m)}$

} **until** $\frac{\frac{1}{m} \sum_{i=1}^m \|x^{(i)} - x_{approx}^{(i)}\|^2}{\frac{1}{m} \sum_{i=1}^m \|x^{(i)}\|^2} \leq 0.01$

انتفاپ تعداد مؤلفه‌های اصلی

۲۶

```
m, n = X.shape
X = X - X.mean(axis=0)
Sigma = (X.T @ X) / m
U, S, V = np.linalg.svd(Sigma)

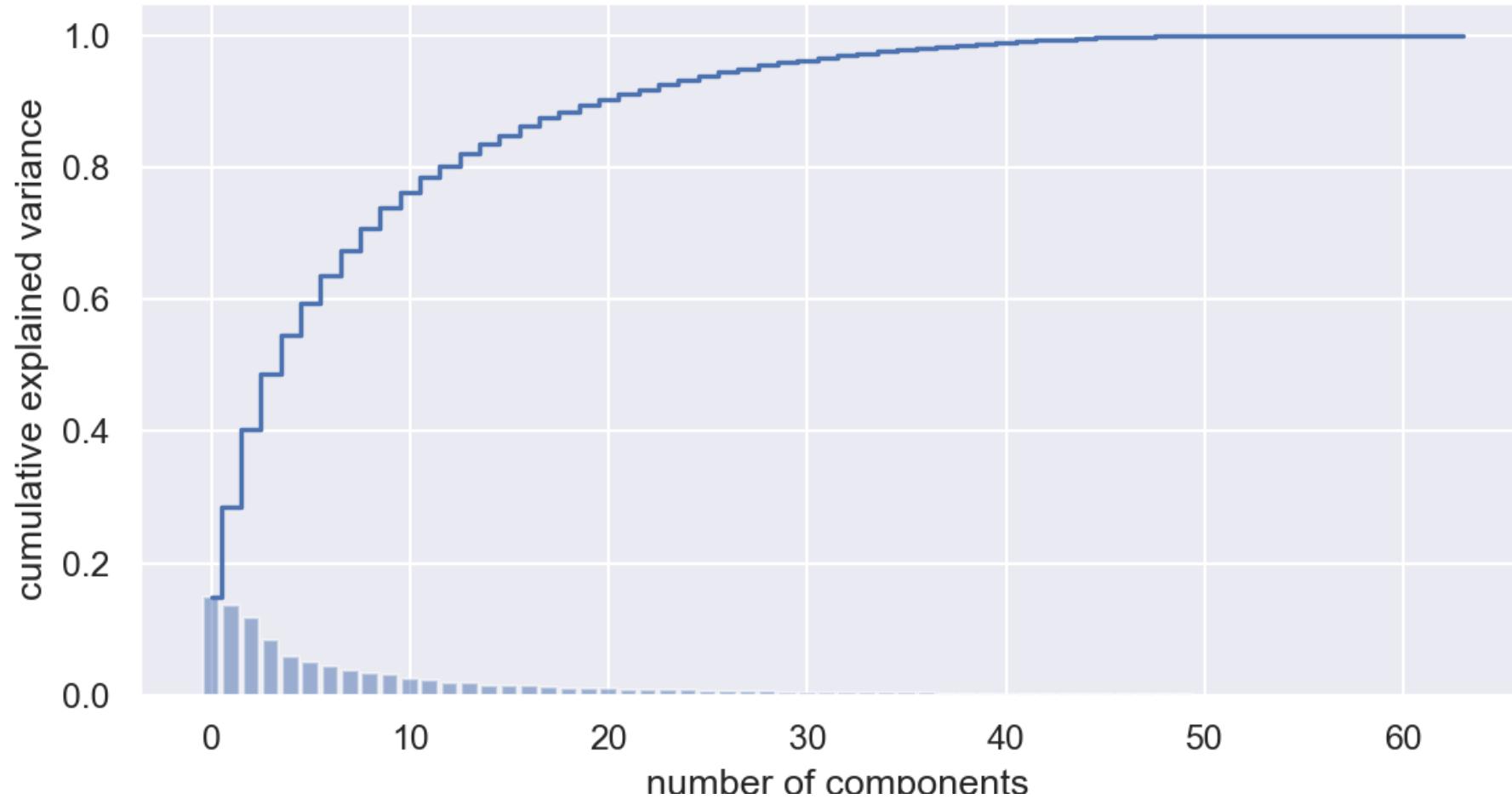
for k in range(1, n + 1):
    total_var = np.sum(S[:k]) / np.sum(S)
    if total_var >= 0.99: break
return k
```

$$S = \begin{bmatrix} s_{11} & 0 & \cdots & 0 \\ 0 & s_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_{nn} \end{bmatrix}$$

[8.72e+00 1.58e+00 6.69e-02 4.79e-16 1.35e-47]

انتساب تعداد مؤلفه‌های اصلی: ارقام ۸ در ۸

۲۷



انتفاپ تعداد مؤلفه‌های اصلی

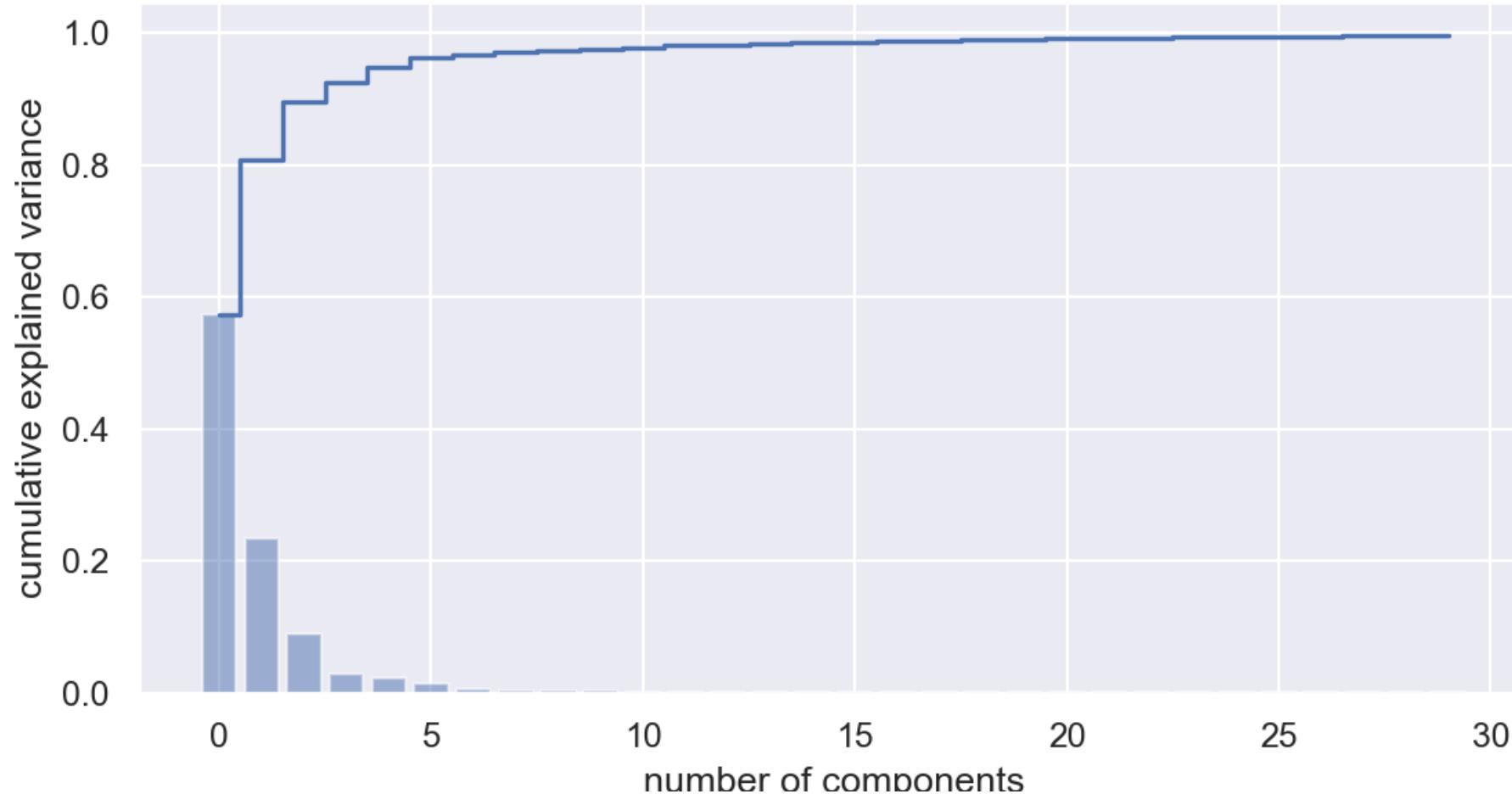
۲۸

داده‌های مربوط به نیمه‌هادی‌ها [۵۹۰ ویژگی] □

تعداد مؤلفه‌ها	درصد واریانس	درصد تجمعی
۱	۵۹/۲	۵۹/۲
۲	۲۴/۱	۸۳/۴
۳	۹/۲	۹۲/۵
۴	۲/۳	۹۴/۸
۵	۱/۵	۹۶/۳
۶	۰/۵	۹۶/۸
۷	۰/۳	۹۷/۱
۲۰	۰/۰۸	۹۹/۳

انتساب تعداد مولفه‌های اصلی

۲۹



استفاده نادرست از PCA: مقابله با بیشبرازش

۳۰

روش نادرست.

- استفاده از $(i)^{(i)} Z$ به جای $x^{(i)}$ باعث کاهش تعداد ویژگی‌ها از n به k می‌شود؛
- در نتیجه، با داشتن ویژگی‌های کمتر، احتمال بیشبرازش کاهش می‌یابد.

روش درست.

- در هنگاه کاهش ابعاد، اطلاعات مربوط به خروجی را استفاده نمی‌کند.
- برای مقابله با بیشبرازش از تنظیم استفاده کنید.

$$\min_{\theta} \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m} \sum_{j=1}^n \theta_j^2$$

سفن آفر

۳۱

□ طراحی یک سیستم یادگیری ماشین.

- ایجاد مجموعه آموزشی به صورت $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$
- استفاده از PCA به منظور کاهش ابعاد $x^{(i)}$ ها و به دست آوردن $z^{(i)}$ ها
- اجرای مرحله آموزش بر روی $\{(z^{(1)}, y^{(1)}), (z^{(2)}, y^{(2)}), \dots, (z^{(m)}, y^{(m)})\}$
- آزمایش فرضیه با استفاده از مجموعه آزمایشی: نگاشت $x_{test}^{(i)}$ به $z_{test}^{(i)}$ و محاسبه $h_\theta(z)$ برای هر یک از داده‌های مجموعه آزمایشی.

- یک پرسش مهم. اگر فرآیند فوق را بدون استفاده از PCA انجام دهیم چه می‌شود؟
- همواره ابتدا فرآیند بالا را بدون استفاده از PCA انجام دهید.
 - اگر به پاسخ مطلوب نرسیدیم، آنگاه استفاده از PCA را آزمایش کنید.

کاربرد

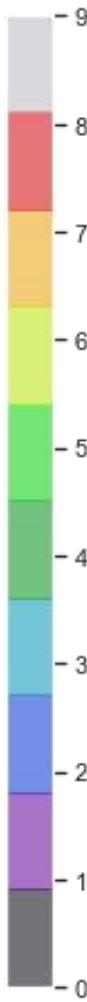
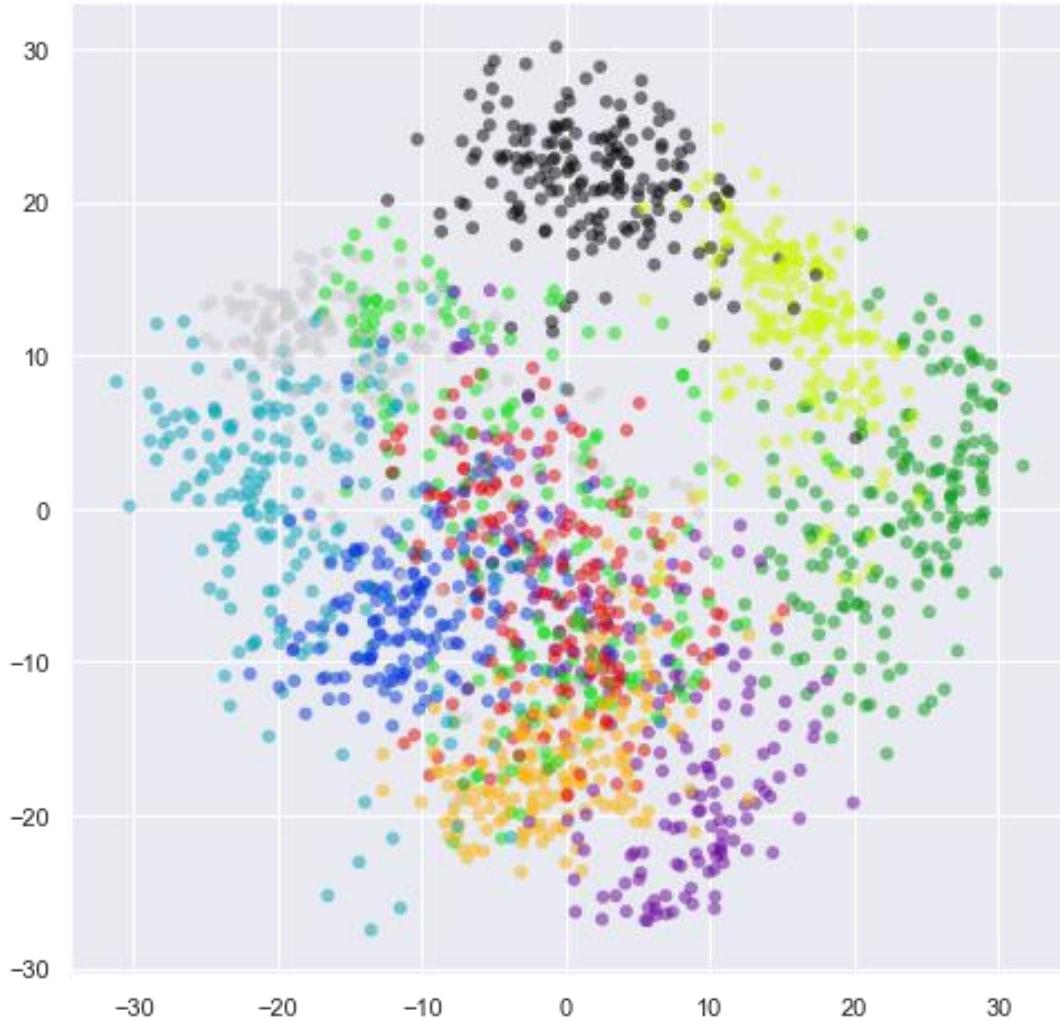
کاربردهای PCA

۳۳

- به تصویر کشیدن داده‌ها.
- انتخاب تعداد مؤلفه‌ها: $k = 3$ یا $k = 2$
- فشرده‌سازی داده‌ها.
- کاهش حافظه مورد نیاز برای ذخیره‌سازی داده‌ها
- افزایش سرعت اجرای الگوریتم یادگیری
- انتخاب تعداد مؤلفه‌ها: بر اساس درصد واریانس حفظ شده

کاربردها: به تصویر کشیدن داده‌ها

۳۴



نگاشت داده‌ها از فضای ۶۴ بعدی به فضای ۲ بعدی
به منظور به تصویر کشیدن و درک بهتر داده‌ها

کاربردها: فشرده‌سازی

۳۵

$k = 100$, variance = 0.91

5041921314
3536172869
4091124327
3869056076
1819398593
3074980941
4460456100
1716302117
8026783904
6746807831

Original Data

5041921314
3536172869
4091124327
3869056076
1819398593
3074980941
4460456100
1716302117
8026783904
6746807831

کاربردها: فشرده‌سازی

۳۶

$k = 150$, variance = 0.95

5041921314
3536172869
4091124327
3869056076
1819398593
3074980941
4460456100
1716302117
8026783904
6746807831

Original Data

5041921314
3536172869
4091124327
3869056076
1819398593
3074980941
4460456100
1716302117
8026783904
6746807831

کاربردها: فشرده‌سازی

۳۷

$k = 200$, variance = 0.97

5041921314
3536172869
4091124327
3869056076
1819398593
3074980941
4460456100
1716302117
8026783904
6746807831

Original Data

5041921314
3536172869
4091124327
3869056076
1819398593
3074980941
4460456100
1716302117
8026783904
6746807831

کاربردها: فشردهسازی

۳۸



- مجموعه آموزشی.
- ۴۰۰ تصویر چهره (خاکستری)
- ابعاد تصاویر.
- ۶۴ در ۶۴ پیکسل
- تعداد ویژگی ها.
- ۴۰۹۶ ویژگی

کاربردها: فشردهسازی

۳۹

$k = 50$, variance = 0.87



Original Faces



کاربردها: فشردهسازی

۴۰

$k = 100$, variance = 0.93



Original Faces



کاربردها: فشردهسازی

۴۱

$k = 150$, variance = 0.96



Original Faces



کاربردها: فشردهسازی

۴۲

$k = 200$, variance = 0.98



Original Faces



کاربردها: فشردهسازی (مؤلفه‌های اصلی)

۴۳

$k = 25$, variance = 0.79



$k = 36$, variance = 0.84



کاربردها: فشردهسازی (بازسازی تصاویر)

۴۴

Principal Components



$k = 25$, variance = 0.79



Original face



کاربردها: فشردهسازی (بازسازی تصاویر)

۴۵

Principal Components



$k = 36$, variance = 0.84



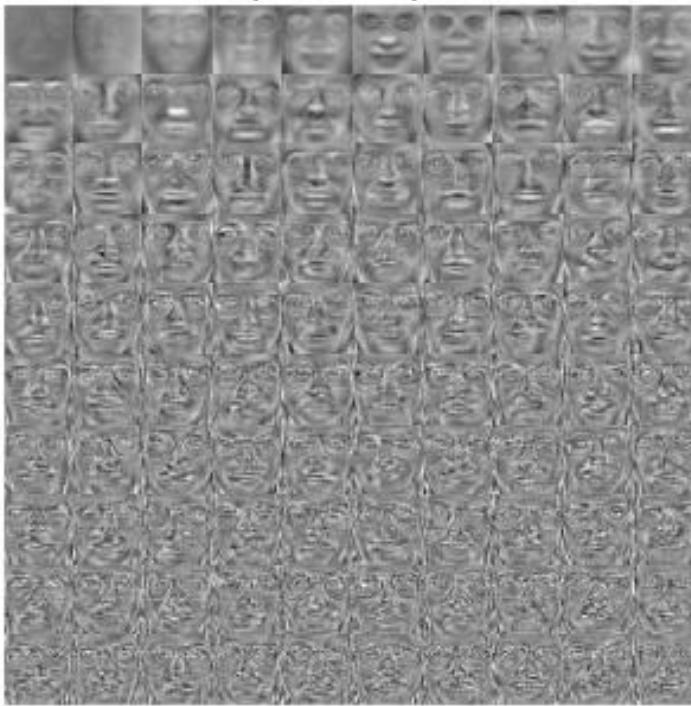
Original face



کاربردها: فشردهسازی (بازسازی تصاویر)

۴۶

Principal Components



$k = 100$, variance = 0.93



Original face



پیوست ا: تجزیه مقادیر منفرد

تجزیه مقادیر منفرد

۴۸

- انگیزه.
- ساده‌سازی داده‌ها
- حذف نویز و افزونگی
- بهبود نتایج الگوریتم
- کاربردهای مثالی.
- جستجو و بازیابی اطلاعات [شاخص گذاری معنایی نهان]
- سیستم‌های توصیه‌گر

تجزیه مقادیر منفرد

۴۹

□ تجزیه مقادیر منفرد.

$$Data_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T$$



ماتریس مقادیر منفرد

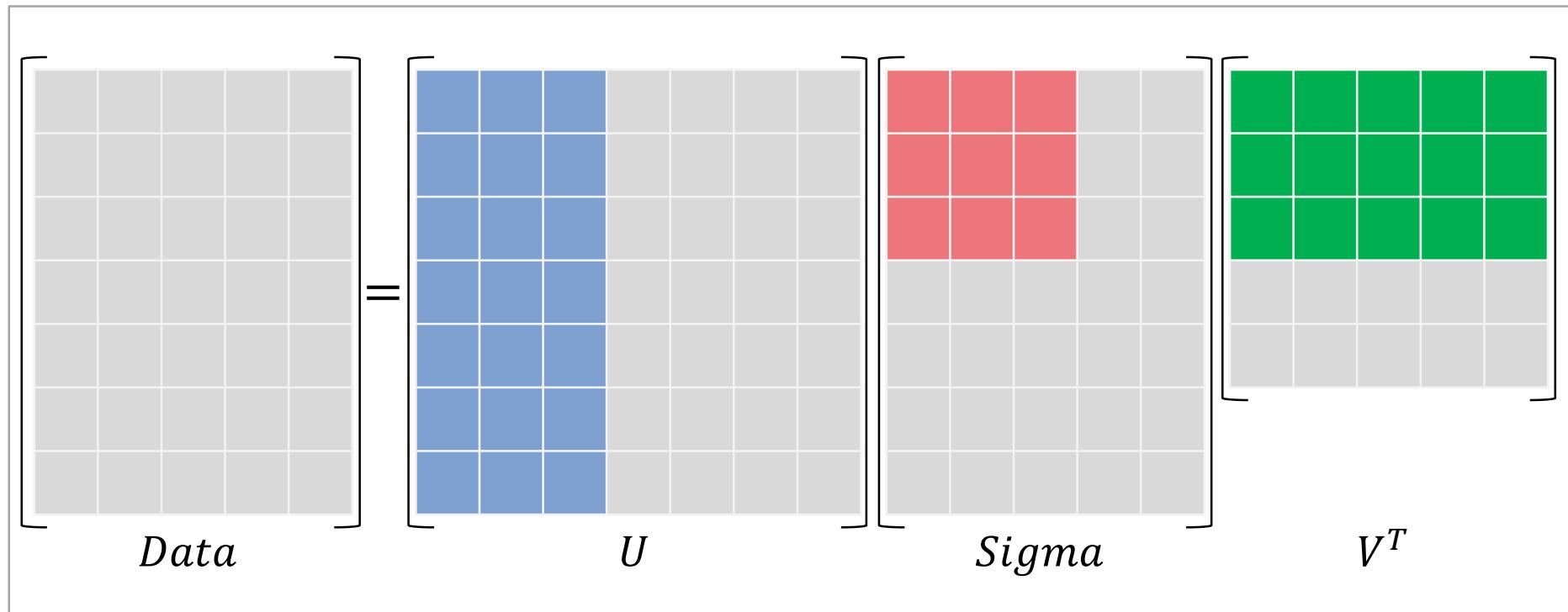
□ ماتریس مقادیر منفرد.

- یک ماتریس قطری که در آن مقادیر منفرد به صورت کاوشی مرتب هستند.
- مقادیر منفرد از یک اندیس مانند σ به بعد دارای مقدار صفر هستند.
- مقادیر منفرد ریشه دوم مقادیر ویژه ماتریس $Data \times Data^T$ هستند.

تجزیه مقادیر منفرد: مثال

۵۰

$$Data_{m \times n} \approx U_{m \times 3} \Sigma_{3 \times 3} V_{3 \times n}$$



تجزیه مقادیر منفرد: مثال

۵۱

```
x = np.array([[1, 1, 1, 0, 0],  
              [2, 2, 2, 0, 0],  
              [1, 1, 1, 0, 0],  
              [5, 5, 5, 0, 0],  
              [1, 1, 0, 2, 2],  
              [0, 0, 0, 3, 3],  
              [0, 0, 0, 1, 1]])
```

```
# Singular Value Decomposition  
U, Sigma, VT = svd(X)  
print(Sigma)
```

[9.72e+00 5.29e+00 6.84e-01 4.12e-16 1.36e-16]

درجه‌های اولیه

1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
1	1	0	2	2
0	0	0	3	3
0	0	0	1	1

تجزیه مقادیر منفرد: مثال

۵۲

```
x_approx = U[:, :1] @ np.diag(Sigma)[:1, :1] @ VT[:1, :]
```

$SSE \approx 28.5$

```
print("SSE = {:.2f}".format(np.linalg.norm(x - x_approx) ** 2))
```

```
x_approx = U[:, :2] @ np.diag(Sigma)[:2, :2] @ VT[:2, :]
```

$SSE \approx 0.47$

```
print("SSE = {:.2f}".format(np.linalg.norm(x - x_approx) ** 2))
```

```
x_approx = U[:, :3] @ np.diag(Sigma)[:3, :3] @ VT[:3, :]
```

$SSE \approx 0.00$

```
print("SSE = {:.2f}".format(np.linalg.norm(x - x_approx) ** 2))
```

تعیین تعداد مقادیر منفرد

۵۳

- تعیین یک تعداد مناسب برای مقادیر منفرد.
- مشابه با تعیین تعداد مؤلفه‌های اصلی
- یک روش تجربی. انتخاب کوچک‌ترین k به طوری که:

$$\frac{\sum_{i=1}^k s_{ii}^2}{\sum_{i=1}^n s_{ii}^2} \geq 0.90$$

$k = 1$. *energy* = 0.768

$k = 2$. *energy* = 0.996

$k = 3$. *energy* = 1.000

9.72	0	0	0	0
0	5.29	0	0	0
0	0	0.68	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

پیوست ۲: مقادیر ویژه و بردارهای ویژه

مقادیر ویژه و بردارهای ویژه

۵۵

□ حاصل ضرب ماتریس مربعی A را در بردار x در نظر بگیرید:

$$Ax = y$$

□ ماتریس A همانند یک تابع بردار x را به بردار جدید y تبدیل می‌کند.

□ بردار ویژه. بردار x یک بردار ویژه است اگر با بردار Ax موازی باشد:

$$Ax = \lambda x$$

مقدار ویژه

مقدار ویژه

□ مقدار ویژه. در رابطه بالا، λ مقدار ویژه متناظر با بردار ویژه x است.

مقادیر ویژه و بردارهای ویژه

۵۶

□ مثال. اگر A یک ماتریس جایگشت به صورت زیر باشد:

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

□ در این صورت:

$$A \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 1 \times \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \lambda = 1$$

متغیر

$$A \begin{bmatrix} -1 \\ 1 \end{bmatrix} = (-1) \times \begin{bmatrix} -1 \\ 1 \end{bmatrix} \Rightarrow \lambda = -1$$

$$\text{trace}(A) = 0 + 0 = 1 + (-1)$$

□ توجه. مجموع مقادیر ویژه با مجموع عناصر قطر اصلی (اثر) برابر است.

محاسبه مقادیر ویژه و بردارهای ویژه

۵۷

□ محاسبه مقادیر ویژه.

$$Ax = \lambda x \Rightarrow (A - \lambda I)x = 0$$

□ بنابراین ماتریس $A - \lambda I$ یک ماتریس منفرد است. [زیرا بردار پوچ دارد]

$$\det(A - \lambda I) = 0$$

□ در نتیجه:

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

$$\det(A - \lambda I) = \det \left(\begin{bmatrix} 3 - \lambda & 1 \\ 1 & 3 - \lambda \end{bmatrix} \right) = \lambda^2 - 6\lambda + 8 = 0 \Rightarrow \lambda = 4, 2$$

trace(A) det(A)

□ مثال. محاسبه مقادیر ویژه

حسابه مقادیر ویژه و بردارهای ویژه

۵۸

□ محاسبه بردارهای ویژه.

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \Rightarrow \lambda_1 = 4, \lambda_2 = 2$$

□ مثال.

$$(A - 4I)x_1 = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}x_1 = 0 \Rightarrow x_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$(A - 2I)x_2 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}x_2 = 0 \Rightarrow x_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

□ مشاهده.

$$Ax = \lambda x \Rightarrow (A + 3I)x = Ax + 3x = \lambda x + 3x = (\lambda + 3)x$$

تجزیه ماتریس A: قطری سازی

۵۹

فرض کنید S ماتریسی باشد که ستون‌های آن بردارهای ویژه ماتریس A هستند. \square

$$\begin{aligned} AS &= A \begin{bmatrix} | & | & \cdots & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & \cdots & | \end{bmatrix} \\ &= \begin{bmatrix} | & | & \cdots & | \\ \lambda_1 x_1 & \lambda_2 x_2 & \cdots & \lambda_n x_n \\ | & | & \cdots & | \end{bmatrix} \\ &= \begin{bmatrix} | & | & \cdots & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \\ &= S\Lambda \end{aligned}$$

$$AS = S\Lambda \Rightarrow S^{-1}AS = \Lambda$$

$$AS = S\Lambda \Rightarrow A = S\Lambda S^{-1}$$

تجزیه ماتریس A : قطری سازی

۶۰

مشاهده. اگر ماتریس A را به توان دو برسانیم، مقادیر ویژه به توان دو می‌رسند و بردارهای ویژه تغییر نمی‌کنند. □

$$A = S\Lambda S^{-1} \Rightarrow A^2 = S\Lambda S^{-1} S\Lambda S^{-1} = S\Lambda^2 S^{-1}$$

$$A = S\Lambda S^{-1} \Rightarrow A^k = S\Lambda^k S^{-1}$$
 به طور کلی. □

تجزیه ماتریس متقارن

۶۱

- در یک ماتریس متقارن حقیقی:
 - مقادیر ویژه **حقیقی** هستند.
 - بردارهای ویژه **متعامد نرمال** هستند.

$$A = Q\Lambda Q^{-1} = Q\Lambda Q^T$$

$$\begin{aligned} A &= \begin{bmatrix} | & | & \cdots & | \\ q_1 & q_2 & \cdots & q_n \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} - & q_1^T & - \\ - & q_2^T & - \\ \vdots & \vdots & \vdots \\ - & q_n^T & - \end{bmatrix} \\ &= \lambda_1 q_1 q_1^T + \lambda_2 q_2 q_2^T + \cdots + \lambda_n q_n q_n^T \end{aligned}$$

- مشاهده. هر ماتریس متقارن ترکیب خطی یک مجموعه از ماتریس‌های پروجکشن متعامد است.

ماتریس‌های مثبت معین متقارن

۶۲

- ماتریس مثبت معین. ماتریس A مثبت معین است اگر به ازای هر بردار غیر صفر مانند x :

$$x^T A x > 0$$

- در یک ماتریس مثبت معین متقارن:
 - مقادیر ویژه همگی مثبت هستند.
 - محورها همگی مثبت هستند.
 - دترمینان‌ها همگی (دترمینان زیرماتریس‌های مقدم) مثبت هستند.

□ مثال.

$$A = \begin{bmatrix} 5 & 2 \\ 2 & 3 \end{bmatrix} \Rightarrow \lambda = 4 \pm \sqrt{5}, p_1 = 5, p_2 = \frac{11}{5}, \det(A) = 11$$

تشفیص آنومالی

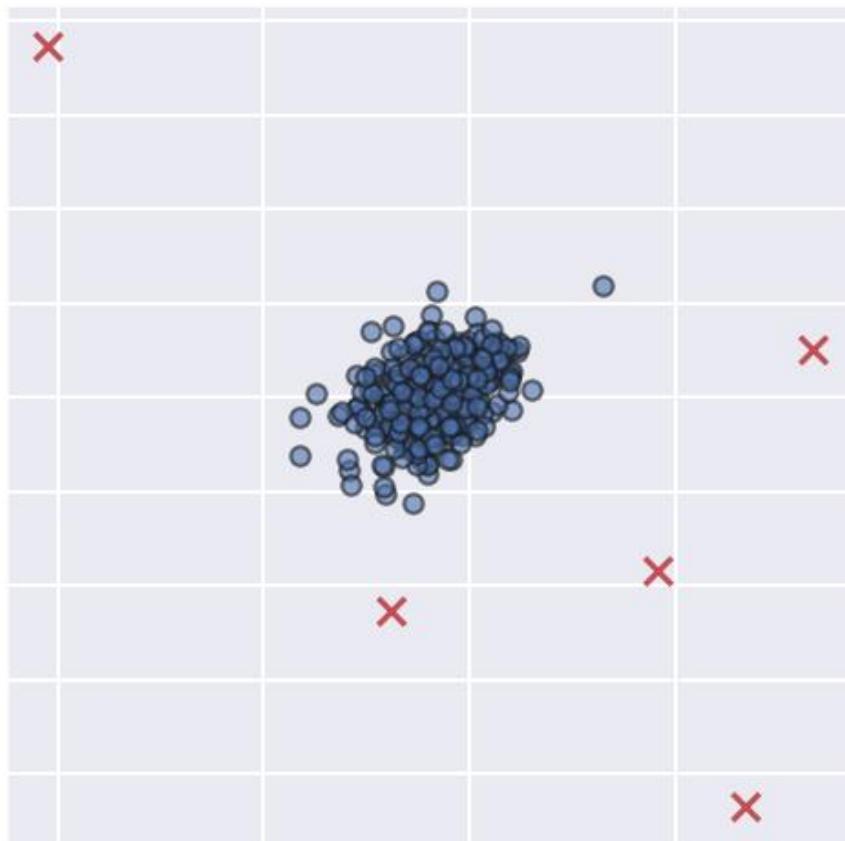
سید ناصر رضوی www.snrazavi.ir

۱۳۹۷

تشخیص آنومالی [تشخیص داده‌های پرت]

۲

- تشخیص آنومالی. تشخیص مشاهداتی که با اغلب مشاهدات انجام شده به شدت تفاوت دارند.

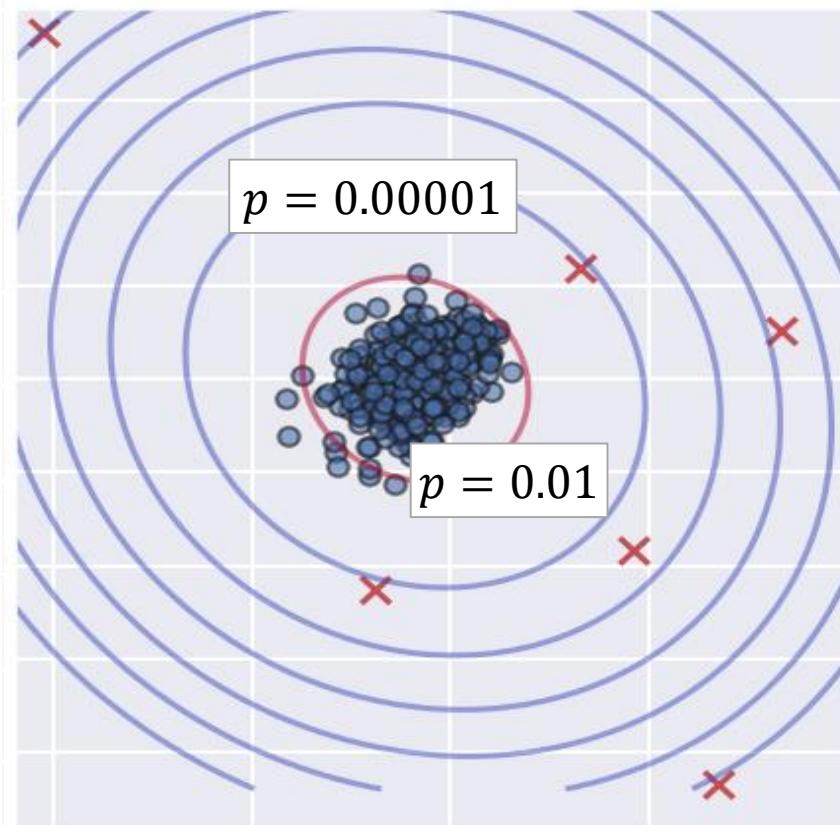


- تشخیص کلاهبرداری.
- تشخیص تراکنش‌های بسیار نامحتمل توسط شخص دارای کارت اعتباری
- امنیت شبکه.
- تشخیص فعالیت‌هایی که با احتمال بسیار کمی به وسیله یک کاربر قانونی انجام شده‌اند.

تشخیص آنومالی [تشخیص داده‌های پرت]

۳

- تشخیص آنومالی. تشخیص مشاهداتی که با اغلب مشاهدات انجام شده به شدت تفاوت دارند.



- رویکرد احتمالاتی برای تشخیص آنومالی.
- ایجاد یک مدل احتمالاتی از داده‌ها
[بیانگر احتمال مشاهده هر رویداد ممکن]
- مشخص کردن مشاهداتی که احتمال وقوع آنها بسیار کم است.

$$p(x) < \epsilon$$

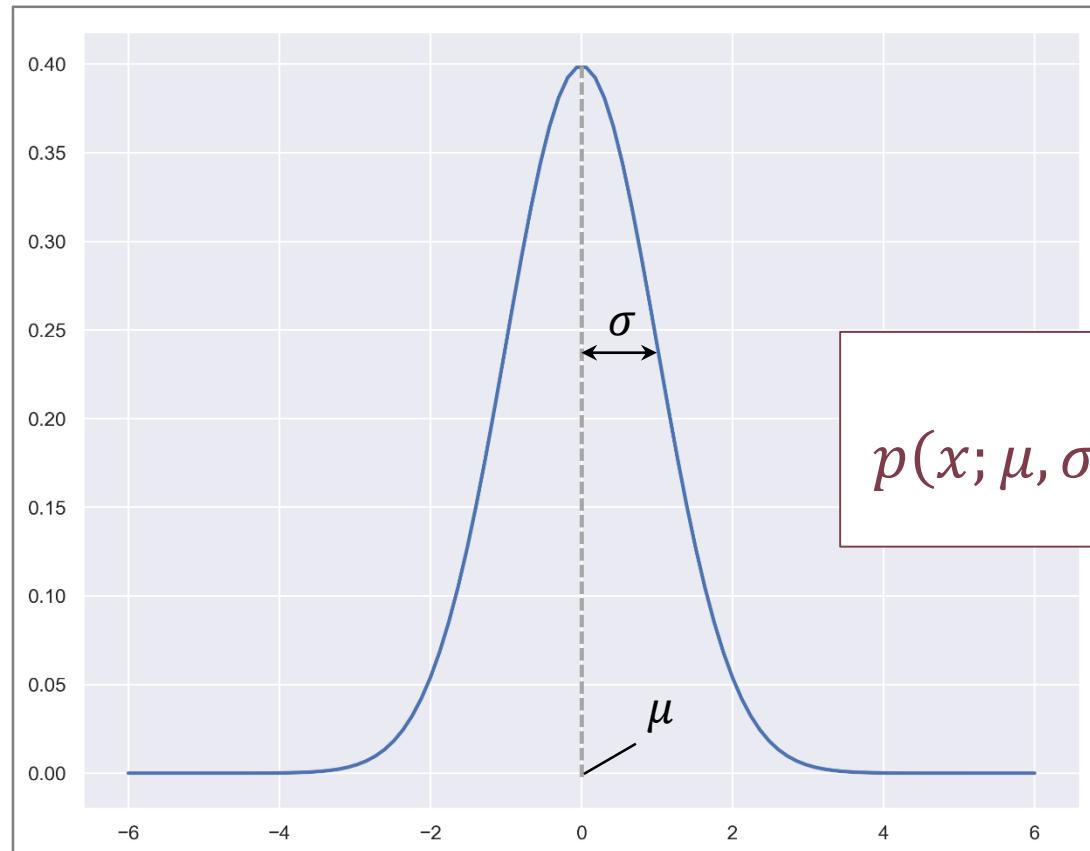
توزيع گوسي (نرمال)

توزیع گوسی

۵

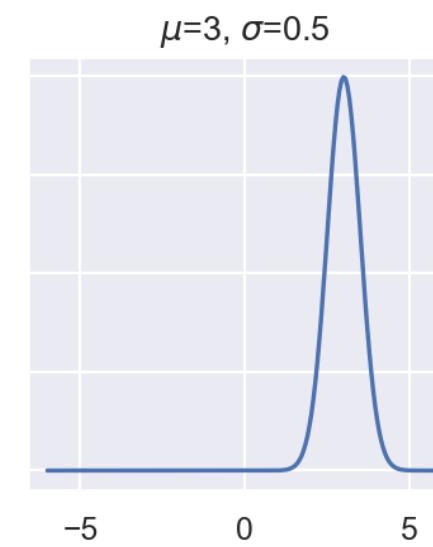
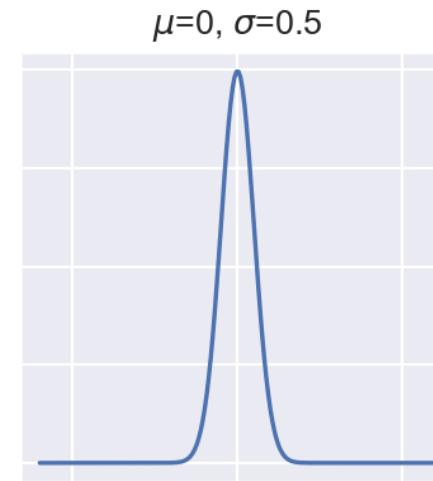
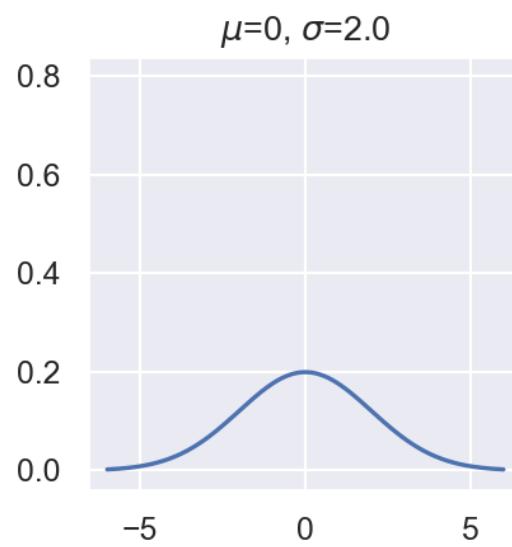
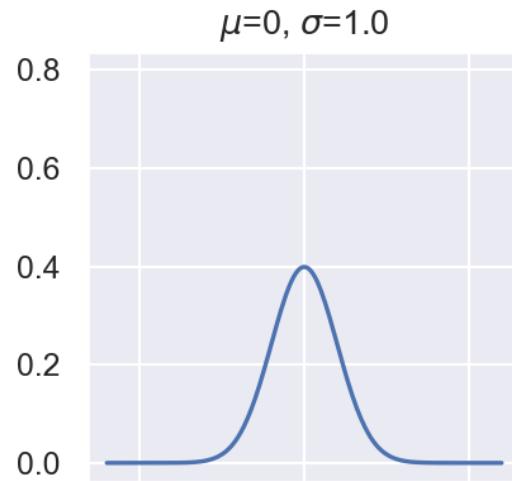
□ توزیع گوسی. فرض کنید x دارای توزیع گوسی با میانگین μ و واریانس σ^2 باشد.

$$x \sim \mathcal{N}(\mu, \sigma^2)$$



توزیع گوسی تک متغیره

۶



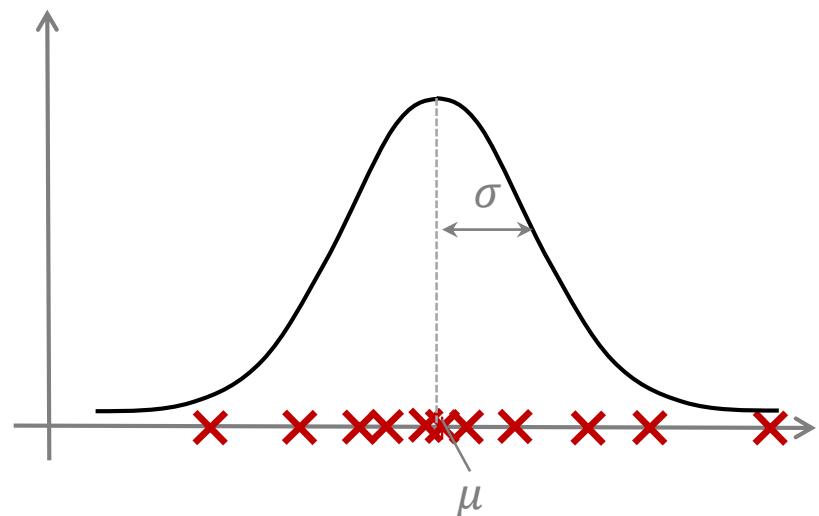
تَفْهِيمِين پارامتر

γ

مجموعه داده. □

$$\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(m)}\}$$

هدف. تخمین مقادیر μ و σ □



$$\mu = \frac{1}{m} \sum_{i=1}^m x^{(i)}$$

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu)^2$$

الگوريتم تشفير آنومالي

الگوریتم تفمین توزیع

۹

$$\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(m)}\}, \quad x^{(i)} \in \mathbb{R}^n$$

□ مجموعه آموزشی.

□ فرضیات.

$$x_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

□ ویژگی‌ها از توزیع نرمال پیروی می‌کنند.

□ بین ویژگی‌ها همبستگی وجود ندارد. [ماتریس کوواریانس قطری است]

$$p(x) = p(x_1; \mu_1, \sigma_1^2)p(x_2; \mu_2, \sigma_2^2)p(x_3; \mu_3, \sigma_3^2) \cdots p(x_n; \mu_n, \sigma_n^2)$$

$$= \prod_{j=1}^n p(x_j; \mu_j, \sigma_j^2)$$

الگوریتم تشخیص آنومالی

۱۰

□ تعیین ویژگی‌هایی که می‌توانند در تشخیص آنومالی مفید باشند.

□ تخمین پارامترها (به ازای $1 \leq j \leq n$)

$$\mu_j = \frac{1}{m} \sum_{i=1}^m x_j^{(i)}$$

$$\sigma_j^2 = \frac{1}{m} \sum_{i=1}^m (x_j^{(i)} - \mu_j)^2$$

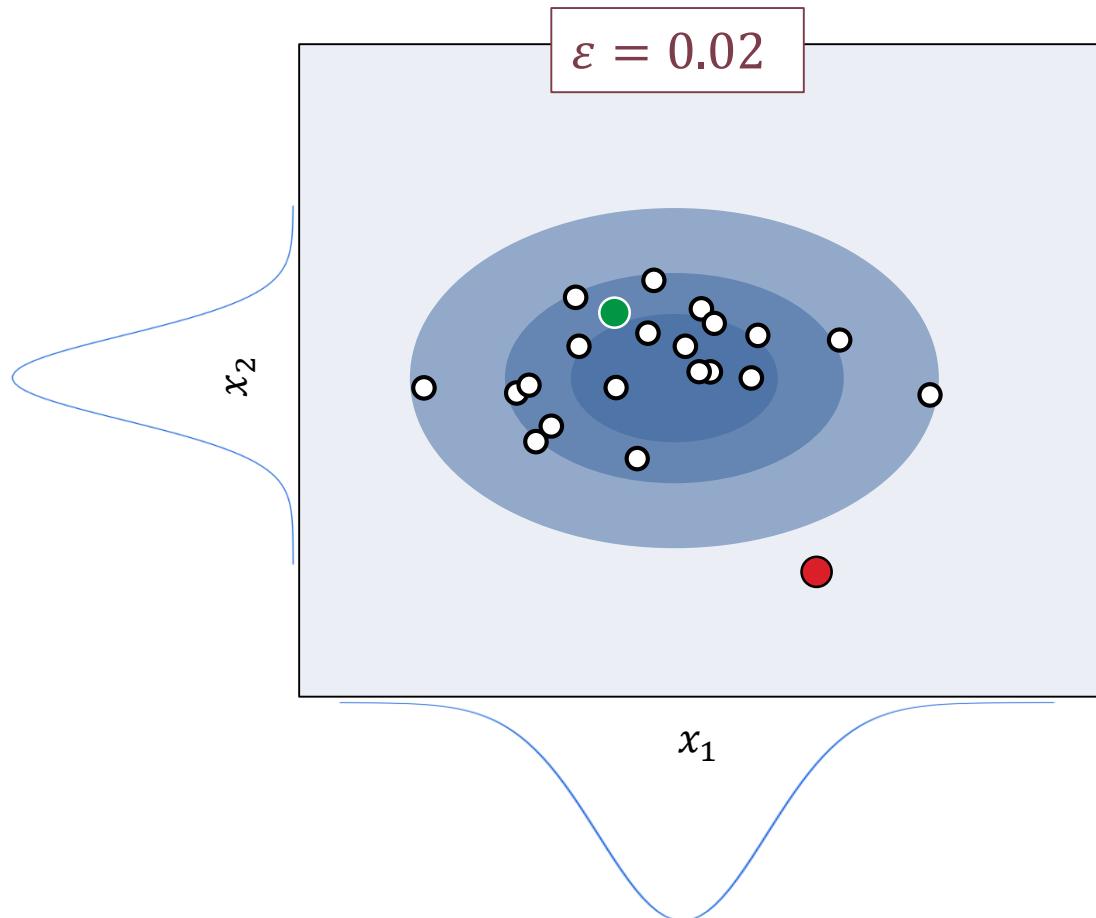
□ محاسبه $p(x)$ برای داده جدید x

$$p(x) = \prod_{j=1}^n p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^n \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right)$$

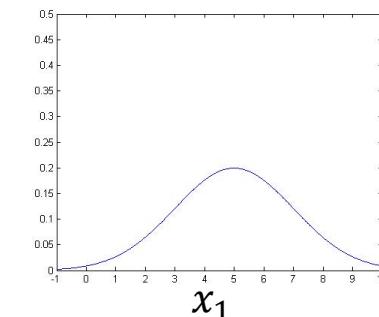
□ تولید خروجی «بله» به شرطی که $p(x) < \epsilon$

مثال

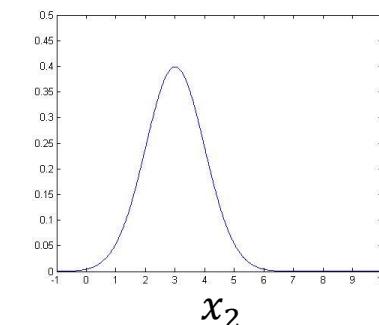
۱۱



$$\mu_1 = 5, \sigma_1 = 2$$



$$\mu_2 = 3, \sigma_2 = 1$$



$$p(x_{test}^{(1)}) = 0.0426$$

$$p(x_{test}^{(2)}) = 0.0021$$

توسعه و ارزیابی سیستم‌های تشخیص آنومالی

ارزیابی‌های عددی

۱۳

□ اهمیت.

- در طول فرآیند توسعه سیستم‌های یادگیری، اگر روشی به منظور ارزیابی سیستم در اختیار داشته باشیم، آنگاه بسیاری از تصمیم‌گیری‌ها (همانند انتخاب ویژگی‌ها و غیره) بسیار ساده‌تر خواهد شد.
- فرض کنید تعدادی داده برچسب‌گذاری شده داریم به طوری که به ازای هر داده، عادی بودن ($y = 0$) و یا غیرعادی بودن ($y = 1$) آن مشخص شده است.

$$\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(m)}\}$$

□ مجموعه آموزشی. [شامل داده‌های عادی]

$$\left\{\left(x_{cv}^{(1)}, y_{cv}^{(1)}\right), \left(x_{cv}^{(2)}, y_{cv}^{(2)}\right), \dots, \left(x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})}\right)\right\}$$

□ مجموعه اعتبارسنجی.

$$\left\{\left(x_{test}^{(1)}, y_{test}^{(1)}\right), \left(x_{test}^{(2)}, y_{test}^{(2)}\right), \dots, \left(x_{test}^{(m_{test})}, y_{test}^{(m_{test})}\right)\right\}$$

□ مجموعه آزمایشی.

مثال

۱۴

□ مجموعه داده. اطلاعات مربوط به عملکرد موتورها

□ ۱۰۰۰۰ موتور سالم

□ ۲۰ موتور معیوب

□ تقسیم‌بندی داده‌ها.

□ مجموعه آموزشی:

□ مجموعه اعتبارسنجی:

□ مجموعه آزمایشی.

۶۰۰۰ موتور سالم دسته‌بندی تک‌دسته‌ای

۲۰۰۰ موتور سالم و ۱۰ موتور معیوب

۲۰۰۰ موتور سالم و ۱۰ موتور معیوب

ارزیابی الگوریتم

۱۵

- آموزش. توسعه مدل $p(x)$ با توجه به مجموعه آموزشی
- پیش‌بینی. برای نمونه‌های موجود در مجموعه اعتبارسنجی یا آموزشی

$$y = \begin{cases} 1, & p(x) < \varepsilon \\ 0, & p(x) \geq \varepsilon \end{cases}$$

- معیارهای ارزیابی ممکن.

□ مثبت درست، مثبت غلط، منفی درست، منفی غلط

□ نرخ دقت و نرخ یادآوری

□ امتیاز F_1

- توجه. می‌توان از مجموعه اعتبارسنجی به منظور انتخاب یک مقدار مناسب برای ε استفاده نمود.

معیارهای ارزیابی

۱۶

□ معیارهای ارزیابی. برای داده‌های نامتوازن

		واقعی		
		$y = 1$	$y = 0$	
$y = 1$	TP	FP		
	FN	TN		

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \frac{P \cdot R}{P + R}$$

تشفیص آنومالی یا یادگیری نظرات شده؟

تشخیص آنومالی یا یادگیری نظارت شده؟

۱۸

یادگیری نظارت شده

- تعداد نمونه‌ها.
- تعداد نمونه‌های مثبت و منفی زیاد
- نمونه‌های مثبت.
- تعداد نمونه‌های مثبت برای این که الگوریتم درک درستی از نمونه‌های مثبت پیدا کند، کفايت می‌کند.
- نمونه‌های مثبت جدید به احتمال زیاد شبیه به نمونه‌های مثبتی هستند که الگوریتم قبلاً در طی فرآيند آموزش با آنها مواجه شده است.

تشخیص آنومالی

- تعداد نمونه‌ها.
- نسبت تعداد نمونه‌های مثبت به منفی بسیار کم
- «انواع» بسیار متفاوت از آنومالی‌ها.
- برای هر الگوریتمی، یاد گرفتن آنومالی‌ها از روی تعداد کم نمونه‌های مثبت بسیار دشوار است.
- آنومالی‌های جدید ممکن است هیچ شباهتی به آنومالی‌هایی که قبلاً دیده شده‌اند، نداشته باشند.

تشخیص آنومالی یا یادگیری نظارت شده؟

۱۹

یادگیری نظارت شده

- تشخیص هرزname.
- پیش‌بینی وضعیت آب و هوا.
- تشخیص غده‌های سرطانی بدخیم.

... □

تشخیص آنومالی

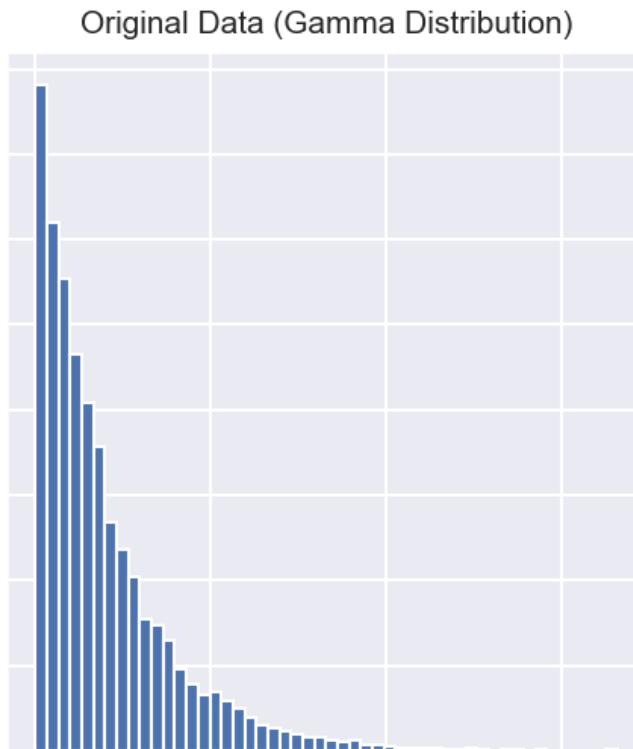
- تشخیص کلاهبرداری.
- ساخت و تولید (ساختن موتور هواپیما).
- نظارت بر ماشین‌ها در مراکز داده‌ای.

... □

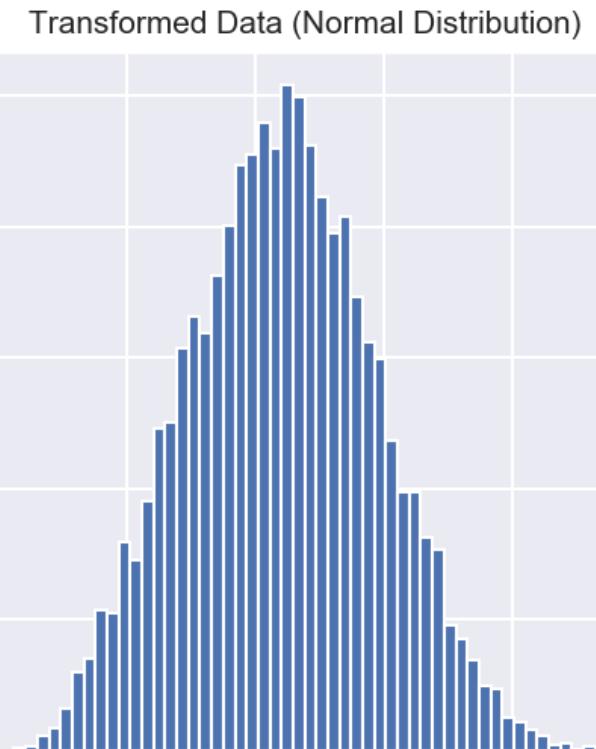
انفاب ویزگی‌ها

تبديل ويلكى با توزيع غيرنرمال به ويلكى با توزيع نرمال

۲۱



$$x^{0.3}$$

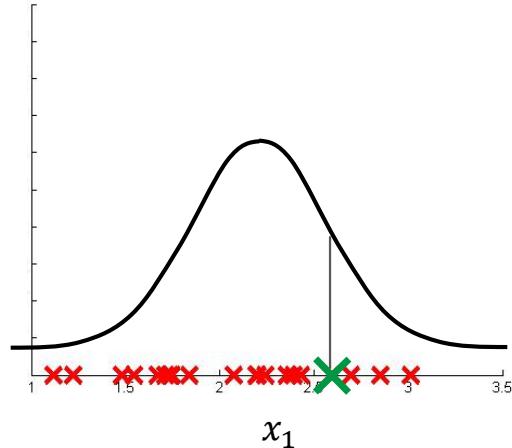


```
x = np.random.gamma(1, 2, (10000, 1))
plt.hist(x, 50)
```

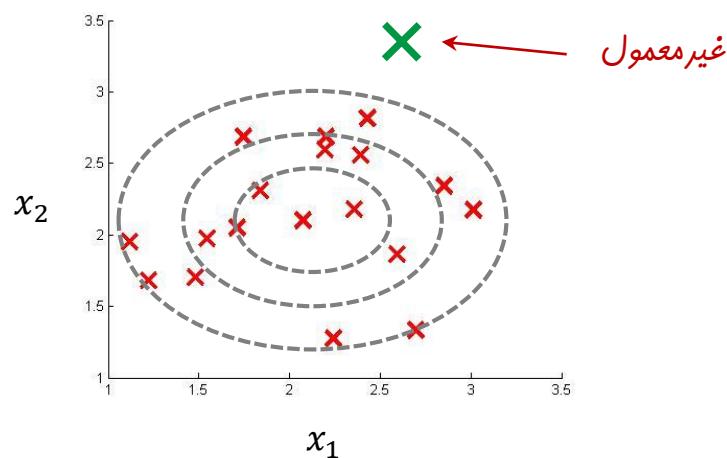
```
plt.hist(x ** 0.3, 50)
```

تملیل فطا برای کمک به تشخیص آنومالی

۲۲



- هدف. می خواهیم مقدار $(x : p(x))$ برای داده های عادی بزرگ باشد.
- برای داده های غیر عادی کوچک باشد.



- یک مشکل متداول.
- $p(x)$ برای داده های عادی و غیر عادی تفاوت چندانی ندارد.

نظارت بر کامپیوٹرها در مرکز داده‌ای

۲۳

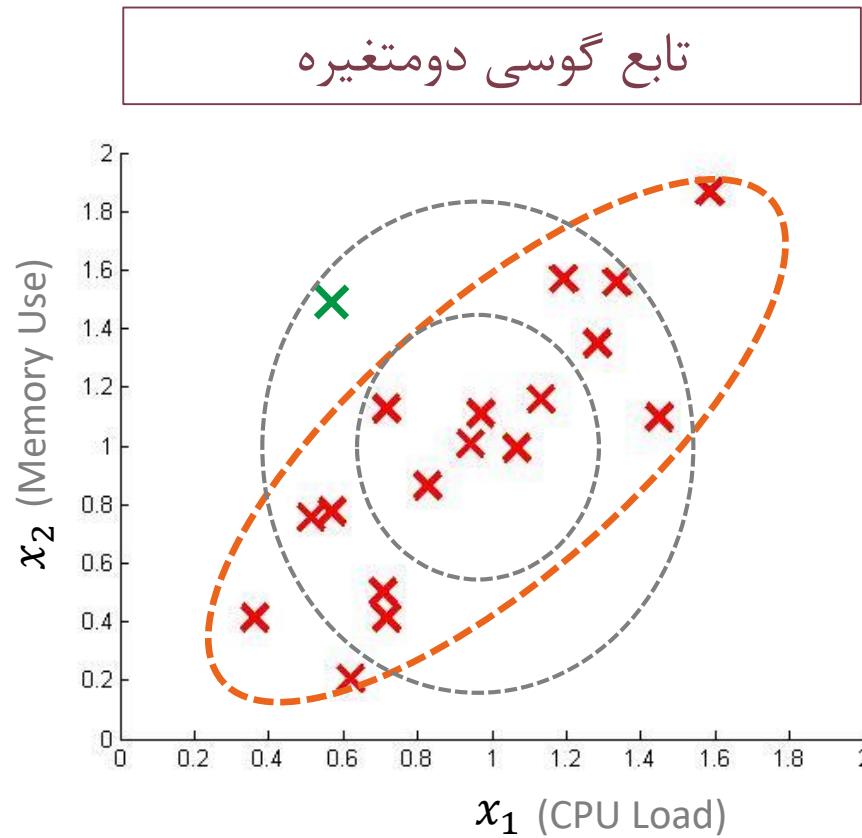
- انتخاب ویژگی‌ها. انتخاب ویژگی‌هایی که در صورت وجود آنومالی مقدارشان بسیار کوچک یا بسیار بزرگ باشد.
 - میزان مصرف حافظه
 - تعداد دستیابی‌ها به دیسک در ثانیه
 - میزان بار پردازنده
 - ترافیک شبکه
- افزودن ویژگی‌های جدید برای تشخیص شرایط غیرعادی.
 - نسبت بار پردازنده به ترافیک شبکه

[مثلاً اگر پردازنده در یک حلقه بینهایت گیر کرده باشد، مقدار این ویژگی بسیار بزرگ خواهد بود]

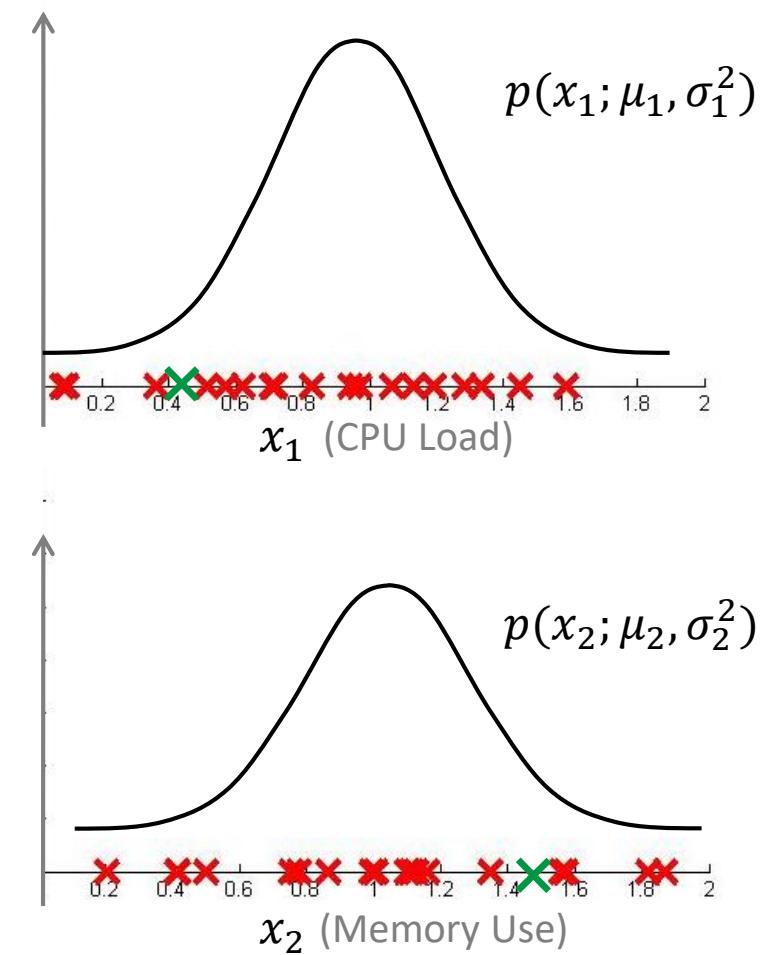
توزيع گوسي پندامنځيره

مثال مقدماتی

۲۵



با افزایش بار پردازندۀ، مصرف حافظه به طور معمول
بیشتر می‌شود. (وجود همبستگی میان ویژگی‌ها)



تابع گوسی چند متغیره

۲۶

□ تابع گوسی چند متغیره.

$$p(x; \mu, \Sigma) = \frac{1}{|\Sigma|^{1/2} (2\pi)^{n/2}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

□ پارامترها.

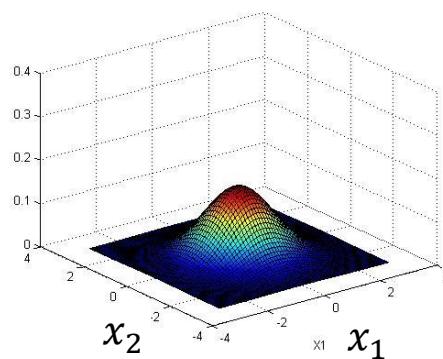
$$\mu \in \mathbb{R}^n$$

ماتریس کوواریانس
↓
 $\Sigma \in \mathbb{R}^{n \times n}$

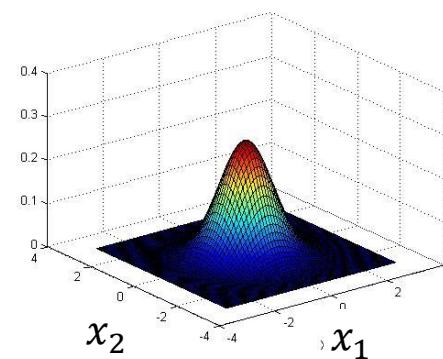
ماتریس کوواریانس قطری، واریانس ویژگی‌ها برابر

۲۷

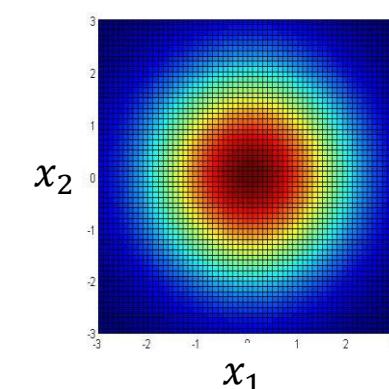
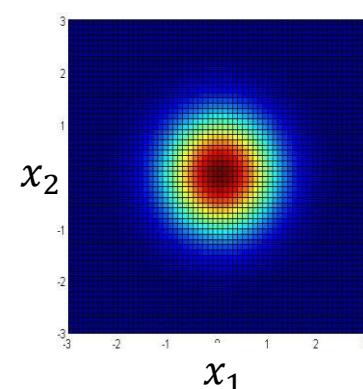
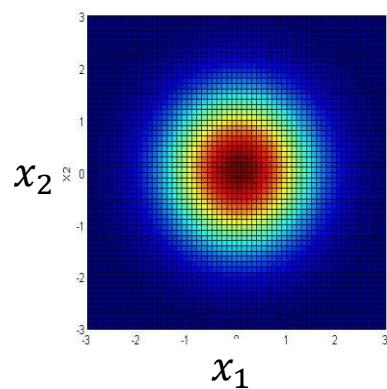
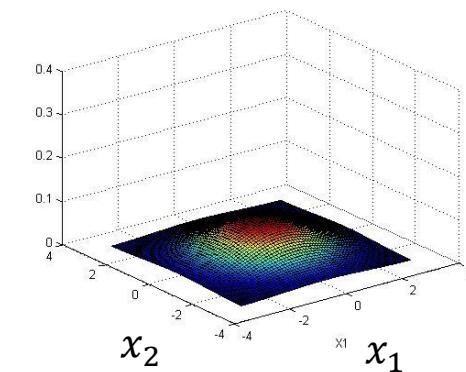
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$$



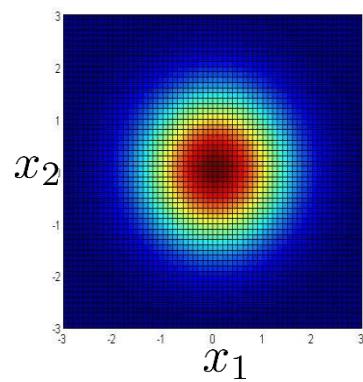
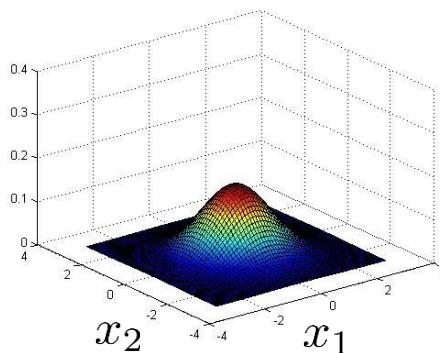
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$



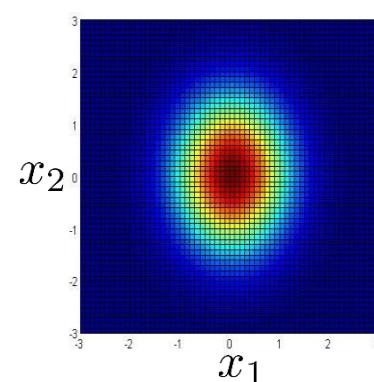
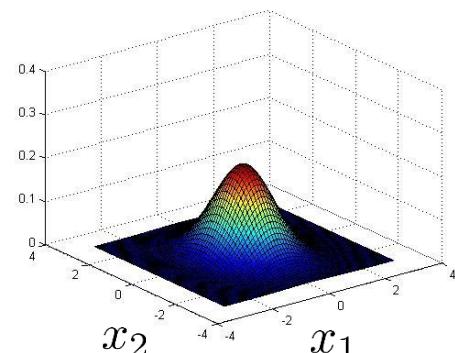
ماتریس کوواریانس قطری، واریانس ویژگی‌ها متفاوت

۲۸

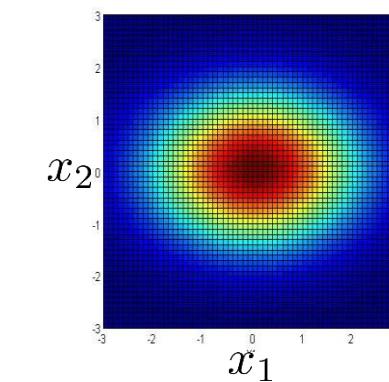
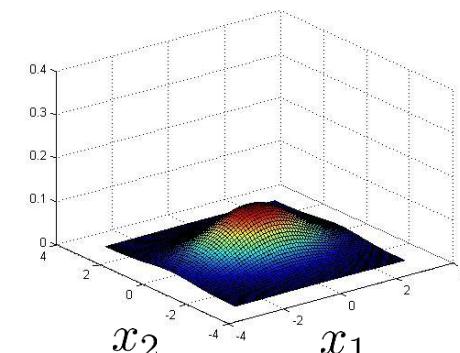
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$$



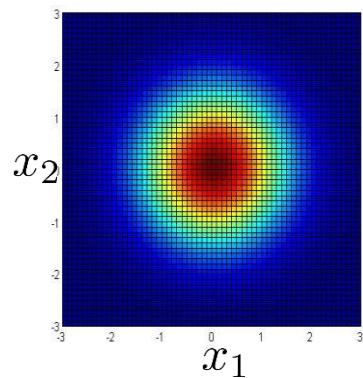
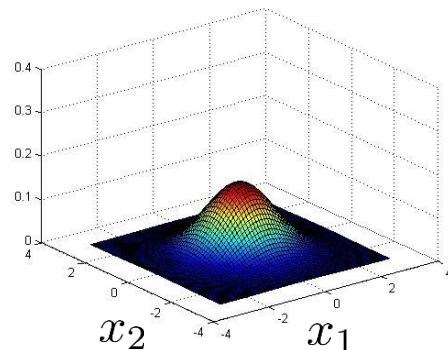
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$



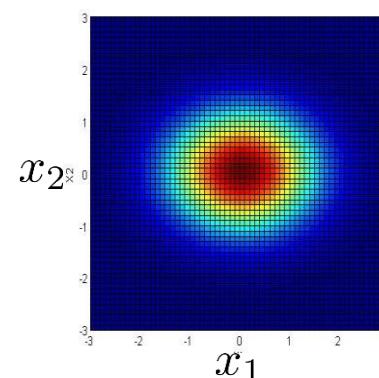
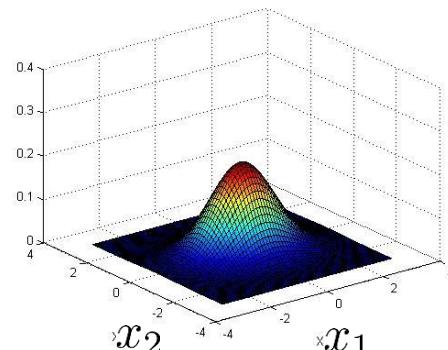
ماتریس کوواریانس قطری، واریانس ویژگی‌ها متفاوت

۲۹

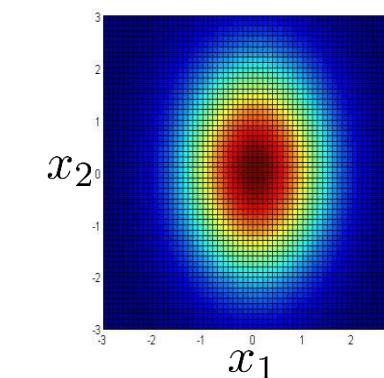
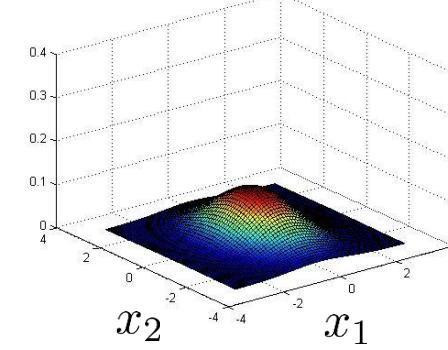
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 0.6 \end{bmatrix}$$



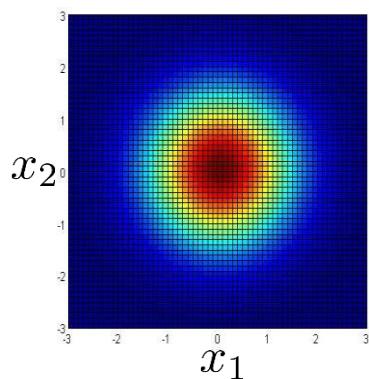
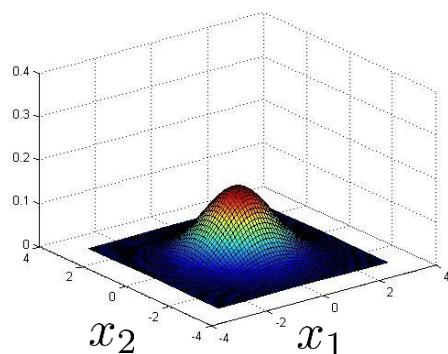
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$



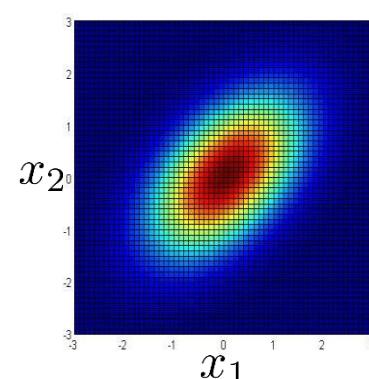
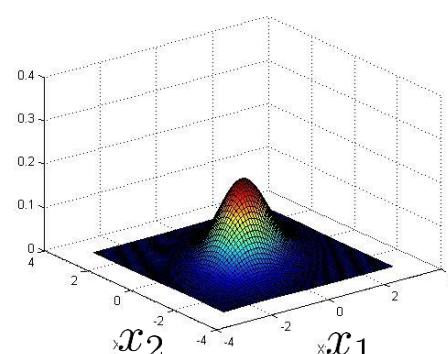
وجود همبستگی مثبت میان ویژگی‌ها

۳۰

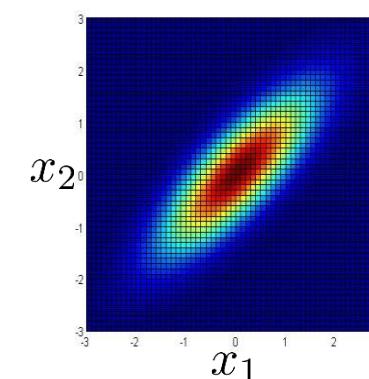
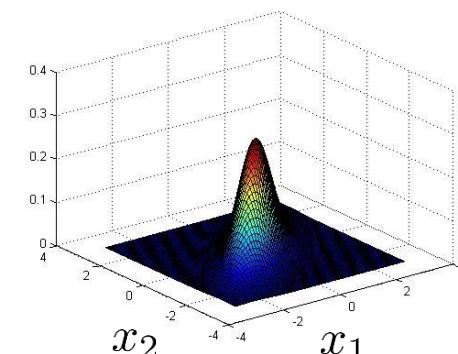
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



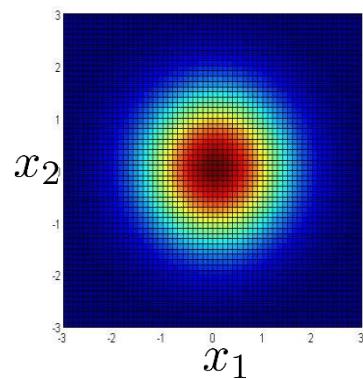
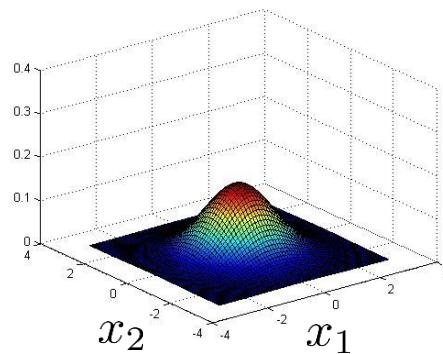
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



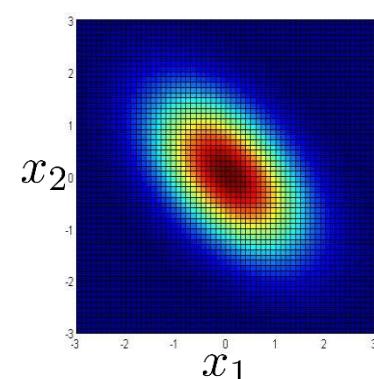
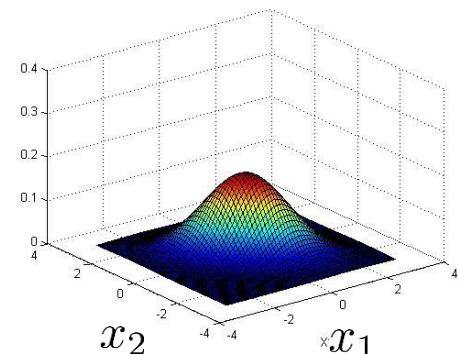
وچود همبستگی منفی میان ویژگی‌ها

۳۱

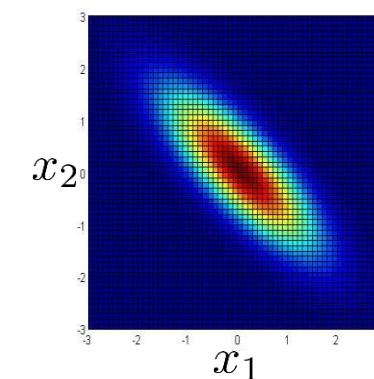
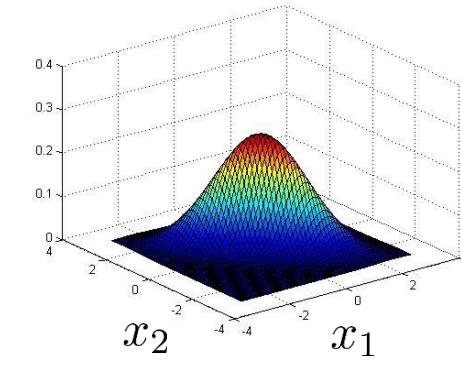
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$



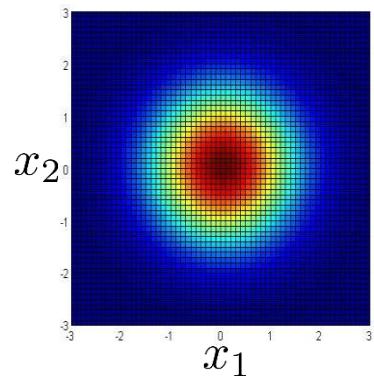
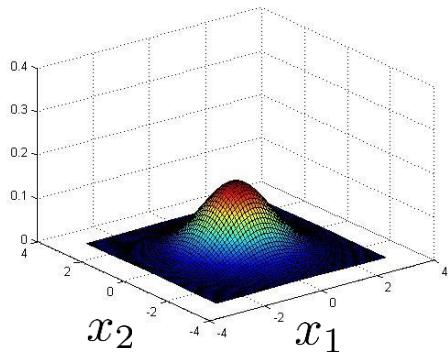
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$



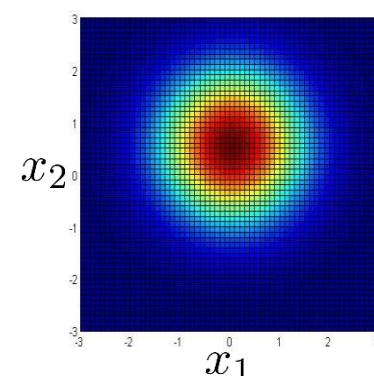
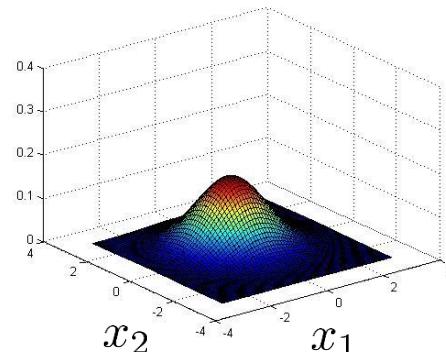
مرکز (میانگین) توزیع گوسی

۳۲

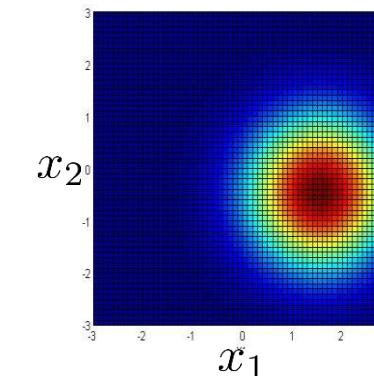
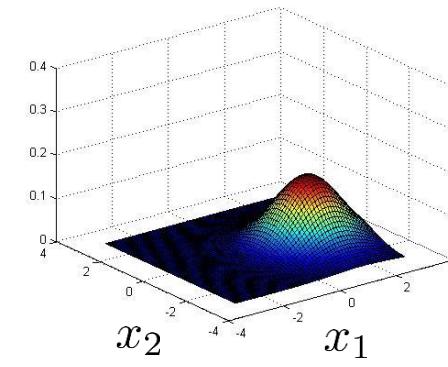
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



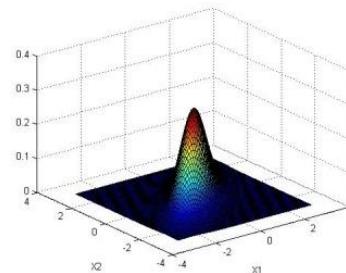
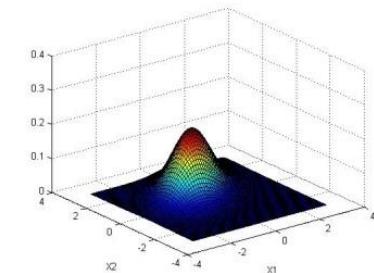
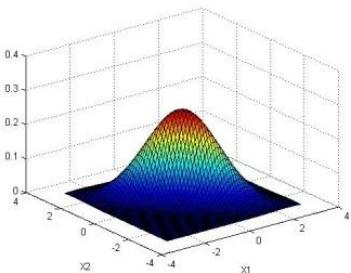
تسلیفیص آنومالی با تابع گوسی پند متخیله

توزيع گوسی چند متغیره

۳۴

□ تابع توزيع گوسی چند متغيره.

$$p(x; \mu, \Sigma) = \frac{1}{|\Sigma|^{1/2} (2\pi)^{n/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$



□ تخمین پارامترها.

$$\mu = \frac{1}{m} \sum_{i=1}^m x^{(i)}$$

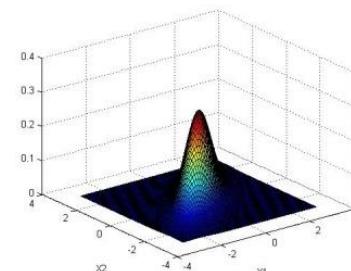
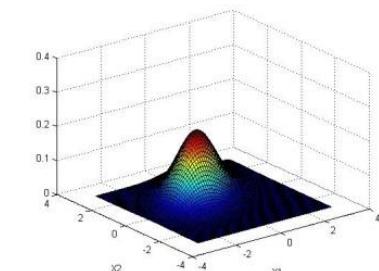
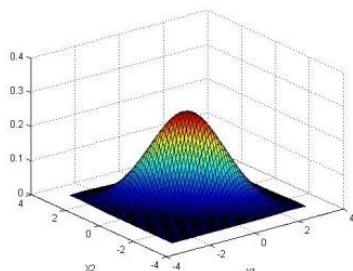
$$\Sigma = \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

توزيع گوسی چند متغیره

۳۵

□ تابع توزيع گوسی چند متغيره.

$$p(x; \mu, \Sigma) = \frac{1}{|\Sigma|^{1/2} (2\pi)^{n/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$



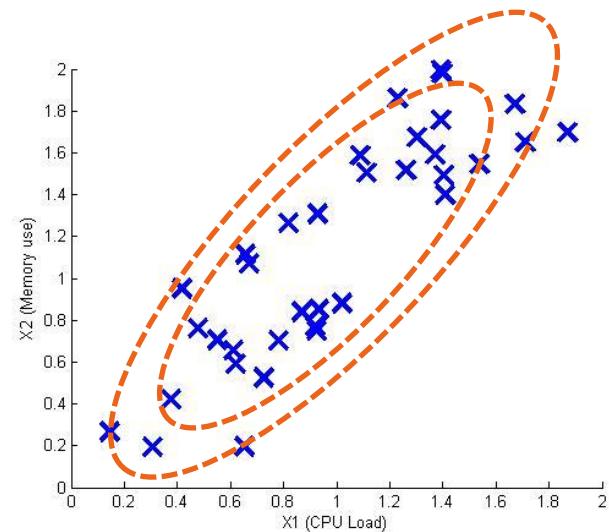
□ تخمین پارامترها.

```
mu = np.mean(X, axis=0)
```

```
Sigma = np.cov(X.T)
```

الگوریتم

۳۶



□ تخمین پارامترهای مدل $p(x)$

$$\mu = \frac{1}{m} \sum_{i=1}^m x^{(i)}$$

$$\Sigma = \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

□ محاسبه مقدار $p(x)$ برای داده‌ی جدید x

$$p(x; \mu, \Sigma) = \frac{1}{|\Sigma|^{1/2} (2\pi)^{n/2}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

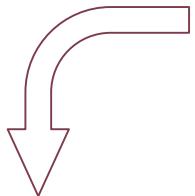
□ تولید خروجی «بله» «اگر $p(x) < \varepsilon$

رابطه با مدل اولیه

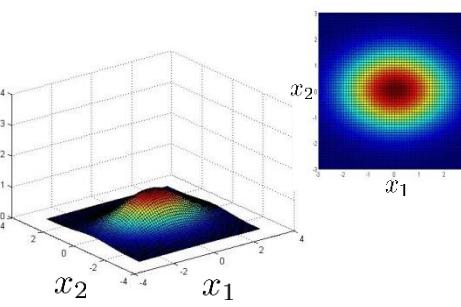
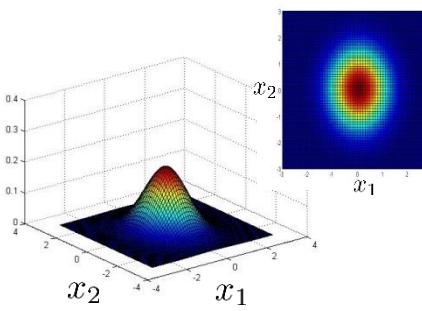
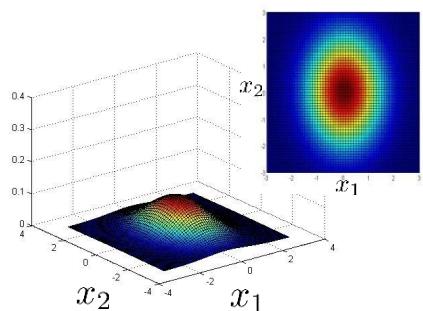
۳۷

□ مدل اولیه.

$$p(x) = p(x_1; \mu_1, \sigma_1^2)p(x_2; \mu_2, \sigma_2^2)p(x_3; \mu_3, \sigma_3^2) \cdots p(x_n; \mu_n, \sigma_n^2)$$



$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$



□ رابطه با توزيع گوسی چند متغيره.

$$p(x; \mu, \Sigma) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

مدل اولیه یا توزیع گوسی چند متغیره

۳۸

□ مدل اولیه.

- ایجاد ویژگی‌ها به صورت دستی انجام می‌شود. (x_1/x_2)
- هزینه‌های محاسباتی به نسبت پایین است.
- اگر تعداد نمونه‌های آموزشی کم باشد، باز هم به درستی عمل می‌کند. [تعداد پارامترها: $2n$]

□ توزیع گوسی چند متغیره.

- به طور خودکار همبستگی میان ویژگی‌ها را یاد می‌گیرد.
- هزینه‌های محاسباتی بالا است. [محاسبه وارون ماتریس کوواریانس]
- تعداد نمونه‌های آموزشی باید از تعداد ویژگی‌ها بیشتر باشد. [وارون‌پذیری ماتریس Σ]

سیستم‌های توصیه‌گر

سید ناصر رضوی www.snrazavi.ir

۱۳۹۷

فهرست

۲

Customers Who Viewed This Item Also Bought

The screenshot shows three recommended items:

- The Diamond Sutra** by Red Pine, Paperback, \$13.57. Click to LOOK INSIDE!
- The Heart Sutra** by Red Pine, Paperback, \$10.17. Click to LOOK INSIDE!
- The Lotus Sutra** by Burton Watson, Paperback, \$18.21. Click to LOOK INSIDE!

Similar Artists

The screenshot shows four recommended artists:

- Stanley Clarke & George Duke**
- Victor Wooten**
- Return to Forever**
- S.M.V.**

□ معرفی

□ رویکردها

□ پالایش گروهی

□ رویکرد مبتنی بر محتوی

□ رویکرد پیوندی

پالائس گروہی

پالایش گروهی

۴

- برجسته‌ترین رویکرد استفاده شده در سیستم‌های توصیه‌گر.
 - استفاده شده به وسیله سایت‌های تجاری بسیار بزرگ
 - شامل انواع مختلفی از الگوریتم‌ها
 - قابل استفاده در بسیاری از دامنه‌ها (کتاب، فیلم، موسیقی و ...)
- رویکرد.
- استفاده از «خرد جمعی» برای توصیه کالاها
- ایده.
- کاربران به کالاهای خریداری شده امتیاز می‌دهند. [عمولاً بین ۱ و ۵]
- کاربرانی که در گذشته سلیقه مشابهی داشته‌اند، احتمالاً در آینده نیز دارای سلیقه‌های مشابهی خواهند بود.

پالایش گروهی

۵

□ **ورودی.** یک ماتریس شامل امتیازهای داده شده به وسیله کاربران به کالاهای گوناگون.

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

□ **نوع خروجی.**

- یک پیش‌بینی (عددی) در مورد میزان علاقه کاربر به یک کالای خاص
- یک فهرست پیشنهادی از N کالای برتر

پالایش گروهی مبتنی بر کاربر

۶

□ روش پایه.

- با داشتن یک کاربر فعال مانند آلیس و کالای i که قبلاً به وسیله آلیس دیده نشده است:
- یک مجموعه از کاربران (نزدیک‌ترین همسایه‌ها) پیدا کن که سلیقه مشابهی با آلیس داشته‌اند و قبلاً به کالای i امتیاز داده‌اند.
- میانگین امتیاز داده شده به وسیله این کاربران به کالای i را محاسبه کن.
- از میانگین محاسبه شده به عنوان تخمینی از میزان علاقمندی آلیس به کالای i استفاده کن
- این کار را برای تمام کالاهایی که آلیس به آنها امتیاز نداده است، تکرار کن.
- کالاهای با امتیاز بیشتر را به آلیس پیشنهاد کن.

□ ایده. کاربرانی که در گذشته سلیقه مشابهی داشته‌اند، احتمالاً در آینده نیز دارای سلیقه‌های مشابهی خواهند بود.

پالایش گروهی مبتنی بر کاربر

۷

□ مثال. ورودی:

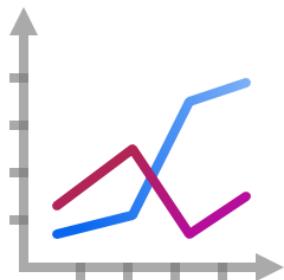
	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

□ هدف. پیش‌بینی میزان علاقمندی آلیس به کالای شماره ۵

پالایش گروهی مبتنی بر کاربر

۸

چند پرسش ابتدایی.



□ چگونه می‌توان شباهت میان کاربران مختلف را محاسبه نمود؟

□ چه تعداد از همسایه‌ها را باید در نظر گرفت؟

□ چگونه می‌توان با توجه به امتیاز همسایه‌ها، یک پیش‌بینی ارائه کرد؟

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

معیارهای اندازه‌گیری شباهت کاربران

۹

ضریب همبستگی پیرسون. یک معیار پر کاربرد
کاربرها : b و a

امتیاز داده شده به وسیله کاربر a به کالای p : $r_{a,p}$
یک مجموعه از کالاها که به وسیله هر دو کاربر a و b امتیازدهی شده‌اند. : P

$$sim(a, b) = \frac{\sum_{p \in P} (r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in P} (r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in P} (r_{b,p} - \bar{r}_b)^2}}$$
$$\frac{cov(a, b)}{std(a) \cdot std(b)}$$

$$\frac{1}{2} + \frac{1}{2} sim(a, b)$$

نرمال‌سازی

معیارهای اندازه گیری شباهت کاربران

۱۰

ضریب همبستگی پیرسون. یک معیار پر کاربرد
کاربرها : b و a :

امتیاز داده شده به وسیله کاربر a به کالای p : $r_{a,p}$

یک مجموعه از کالاها که به وسیله هر دو کاربر a و b امتیازدهی شده‌اند. : P

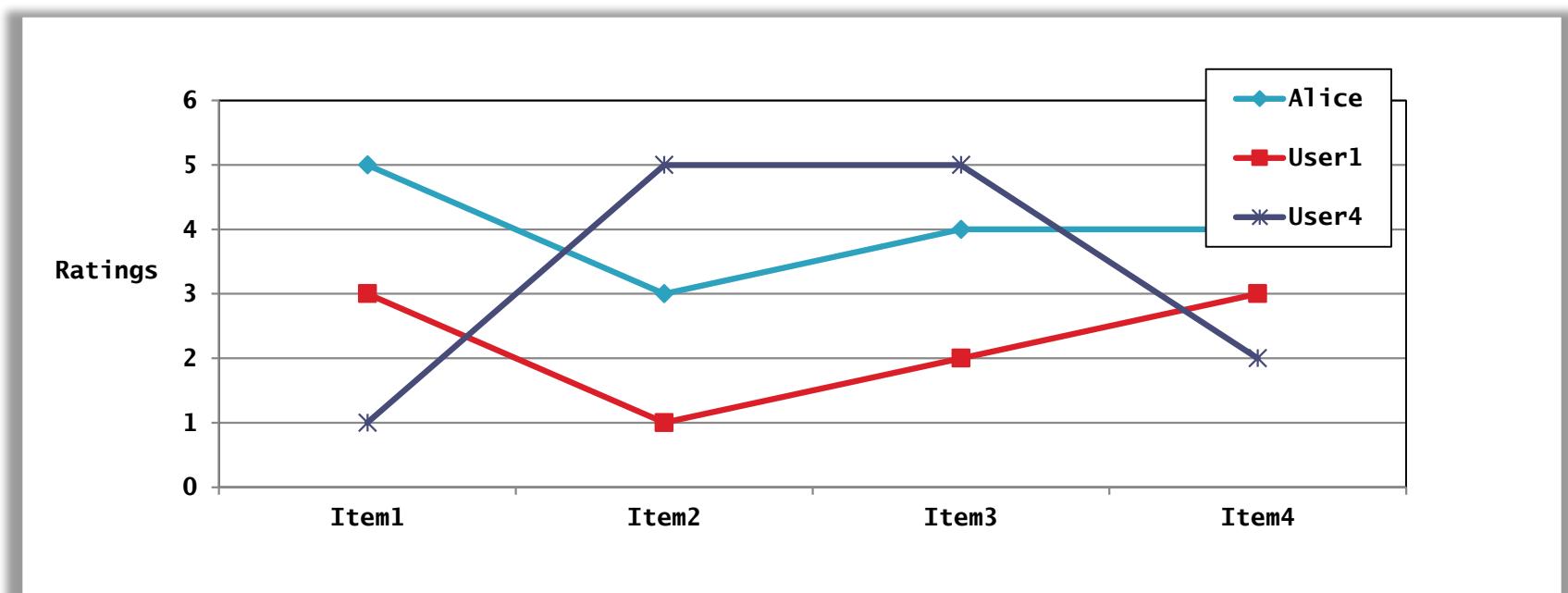
	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

sim = 0.85
sim = 0.70
sim = 0.00
sim = -0.79

ضریب همبستگی پیرسون

۱۱

- مزیت. در نظر گرفتن تفاوت‌ها در عادات امتیازدهی
 - ضریب همبستگی پیرسون در بسیاری از دامنه‌ها نسبت به معیارهای دیگر عملکرد بهتری دارد.
- [معیار اقلیدسی، معیار کسینوسی]



□ یک تابع متداول برای پیش‌بینی:

$$pred(a, p) = \bar{r}_a + \frac{\sum_{b \in N} sim(a, b) * (r_{b,p} - \bar{r}_b)}{\sum_{b \in N} sim(a, b)}$$

- محاسبه این که امتیاز همسایه‌ها برای کالای p کمتر یا بیشتر از میانگین آنها است.
- ترکیب اختلاف‌ها -- استفاده از معیار شباخت برای وزن‌دهی.
- اضافه کردن میانگین امتیاز‌های کاربر a به مقدار محاسبه شده.

(و) یگردهای مبتنی بر حافظه و مبتنی بر مدل

۱۳

□ پالایش گروهی مبتنی بر کاربر، یک روش «مبتنی بر حافظه» است.

□ استفاده مستقیم از ماتریس امتیازها برای یافتن نزدیک‌ترین همسایه‌ها و پیش‌بینی

□ در بسیاری از کاربردهای دنیا واقعی این رویکرد قابل استفاده نیست!

■ به دلیل وجود دهها میلیون کاربر و میلیون‌ها کالا

□ رویکردهای مبتنی بر مدل.

□ بر مبنای یادگیری مدل به صورت آفلاین (آموزش آفلاین)

□ در زمان اجرا تنها از مدل یاد گرفته شده برای پیش‌بینی استفاده می‌شود.

□ مدل‌ها به طور تناوبی به روز رسانی می‌شوند.

□ ایجاد مدل و به روز رسانی آن می‌تواند از نظر محاسباتی بسیار پرهزینه باشد.

پالایش گروهی مبتنی بر کالا

۱۴

ایده.

- ❑ استفاده از شباهت میان کالاها برای پیش‌بینی [نه شباهت میان کاربران]
- ❑ مثال. جستجو به دنبال کالاهای مشابه با کالای ۵
- ❑ استفاده از امتیازهای داده شده به کالاهای مشابه برای پیش‌بینی امتیاز کالای ۵

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

معیار شباهت میان کالاها

۱۵

□ معیار شباهت کسینوسی.

- تولید نتایج بهتر در مقایسه کالا به کالا
- امتیازهای داده شده به هر کالا به عنوان یک بردار در فضای n بعدی در نظر گرفته می‌شوند.
- شباهت میان دو کالا با محاسبه کسینوس زاویه مربوط به بردار این دو کالا اندازه‌گیری می‌شود:

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| * |\vec{b}|}$$

□ معیار شباهت کسینوسی تنظیم شده.

- در نظر گرفتن میانگین امتیازهای هر کاربر
- مجموعه کاربرانی که به هر دو کالای a و b امتیاز داده‌اند.

$$sim(\vec{a}, \vec{b}) = \frac{\sum_{u \in U} (r_{u,a} - \bar{r}_u)(r_{u,b} - \bar{r}_u)}{\sqrt{\sum_{u \in U} (r_{u,a} - \bar{r}_u)^2} \sqrt{\sum_{u \in U} (r_{u,b} - \bar{r}_u)^2}}$$

□ یک تابع متداول برای پیش‌بینی:

$$pred(u, p) = \frac{\sum_{i \in ratedItem(u)} sim(i, p) * r_{u,i}}{\sum_{i \in ratedItem(u)} sim(i, p)}$$

- معمولاً اندازه همسایگی محدود است.
- یعنی، از همه همسایه‌ها برای پیش‌بینی استفاده نمی‌شود.
- یک قاعده تجربی: در بسیاری از کاربردهای دنیای واقعی، تعداد همسایه‌ها بین ۲۰ الی ۵۰ در نظر گرفته می‌شود. [۲۰۰۲]

مسئله خلوت بودن داده‌ها

۱۷

□ مسئله شروع سرد.

□ چگونه می‌توان کالاهای جدید را توصیه کرد؟

□ چگونه می‌توان به کاربران جدید توصیه داد؟

□ راه حل‌های ساده.

□ از کاربر بخواه مجموعه‌ای از کالاها را امتیازدهی کند.

□ در مراحل ابتدایی از روش‌های دیگر مانند **پالایش مبتنی بر محتوی** استفاده کنید.

□ مقادیر پیش‌فرض: استفاده از مقادیر پیش‌فرض برای کالاهایی که فقط یکی از دو کاربری که قرار است مقایسه شوند به آنها امتیاز داده‌اند.

انواع (ویگردهای مبتنی بر مدل

۱۸

- تجزیه ماتریس‌ها.
- تجزیه مقادیر منفرد، تحلیل مولفه‌های اصلی
- کاوش قواعد ارتباطی.
- مقایسه: تحلیل سبد خرید
- مدل‌های احتمالاتی.
- خوشه‌بندی، شبکه‌های بیزی و ...
- هزینه پیش‌پردازش (یادگیری مدل).
- معمولاً درباره آن صحبت نمی‌شود
- آیا به روز رسانی تدریجی ممکن است؟

تجزیه مقادیر منفرد

۱۹

□ انگیزه.

□ ساده‌سازی داده‌ها

□ حذف نویز و افزونگی

□ بهبود نتایج الگوریتم

□ کاربردهای مثالی.

□ جستجو و بازیابی اطلاعات [شاخص‌گذاری معنایی نهان]

□ سیستم‌های توصیه‌گر

تجزیه مقادیر منفرد

۲۰

□ تجزیه مقادیر منفرد.

$$Data_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T$$

ماتریس مقادیر منفرد

□ ماتریس مقادیر منفرد.

- یک ماتریس قطری که در آن مقادیر منفرد به صورت کاهشی مرتب هستند.
- مقادیر منفرد از یک اندیس مانند σ_1 به بعد دارای مقدار صفر هستند.
- مقادیر منفرد ریشه دوم مقادیر ویژه ماتریس $Data \times Data^T$ هستند.

توصیه‌های مبتنی بر محتوی

توصیه مبتنی بر داده‌های

۲۲

آموزش. برای کاربر j بزرگ‌دار $\theta^{(j)} \in \mathbb{R}^3$ را یاد بگیر \square

پیش‌بینی. امتیاز فیلم i برای کاربر j \square

$$(\theta^{(j)})^T x^{(i)}$$

	Alice(1)	Bob(2)	Carol(3)	Dave(4)	x_1	x_2
Titanic	5	5	0	0	0.90	0.00
Sound and Music	5	?	?	0	1.00	0.01
Casablanca	?	4	0	?	0.99	0.00
Fast and Furious	0	0	5	4	0.10	1.00
Desperado	0	0	5	?	0.00	0.90

بیان (لسمنی) مسئله

۲۳

اگر کاربر j به فیلم i امتیاز داده باشد، در غیر این صورت صفر.
 $y^{(i,j)}$ امتیاز داده شده توسط کاربر j به فیلم i

$\theta^{(j)}$ بردار پارامترها برای کاربر j

$x^{(i)}$ بردار ویژگی برای فیلم i

پیش‌بینی امتیاز فیلم i برای کاربر j :

$$(\theta^{(j)})^T x^{(i)}$$

$m^{(j)}$ تعداد فیلم‌های امتیازدهی شده به وسیله کاربر j

هدف بهینه‌سازی

۲۴

□ یادگیری بردار $\theta^{(j)}$ -- پارامترها برای کاربر j

$$\min_{\theta^{(j)}} \frac{1}{2} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{k=1}^n \left(\theta_k^{(j)} \right)^2$$

□ یادگیری بردارهای $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n_u)}$

$$\min_{\theta^{(1)}, \dots, \theta^{(n_u)}} \frac{1}{2} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^n \left(\theta_k^{(j)} \right)^2$$

الگوریتم بهینه‌سازی

۲۵

□ تابع هدف.

$$\min_{\theta^{(1)}, \dots, \theta^{(n_u)}} \frac{1}{2} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^n \left(\theta_k^{(j)} \right)^2$$

□ گرادیان کاہشی.

$$\begin{aligned} \theta_k^{(j)} &= \theta_k^{(j)} - \alpha \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right) x_k^{(i)} && (\text{for } k = 0) \\ \theta_k^{(j)} &= \theta_k^{(j)} - \alpha \left(\sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right) x_k^{(i)} + \lambda \theta_k^{(j)} \right) && (\text{for } k \neq 0) \end{aligned}$$

پالائیش گروہی

پالایش گروهی

۲۷

	Alice(1)	Bob(2)	Carol(3)	Dave(4)	x_1	x_2
Titanic	5	5	0	0	0.90	0.00
Sound and Music	5	?	?	0	1.00	0.01
Casablanca	?	4	0	?	0.99	0.00
Fast and Furious	0	0	5	4	0.10	1.00
Desperado	0	0	5	?	0.00	0.90

$$\theta^{(1)} = \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}$$

$$\theta^{(2)} = \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}$$

$$\theta^{(3)} = \begin{bmatrix} 0 \\ 0 \\ 5 \end{bmatrix}$$

$$\theta^{(4)} = \begin{bmatrix} 0 \\ 0 \\ 5 \end{bmatrix}$$

$$(\theta^{(1)})^T x^{(1)} \approx 5$$

$$(\theta^{(2)})^T x^{(1)} \approx 5$$

$$(\theta^{(3)})^T x^{(1)} \approx 0$$

$$(\theta^{(4)})^T x^{(1)} \approx 0$$

هدف بهینه‌سازی

۲۸

یادگیری $x^{(i)}$ -- با داشتن پارامترهای $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n_u)}$ □

$$\min_{x^{(i)}} \frac{1}{2} \sum_{j:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{k=1}^n \left(x_k^{(i)} \right)^2$$

یادگیری $x^{(1)}, x^{(2)}, \dots, x^{(n_m)}$ -- با داشتن پارامترهای $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n_u)}$ □

$$\min_{x^{(1)}, \dots, x^{(n_m)}} \frac{1}{2} \sum_{i=1}^{n_m} \sum_{j:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{i=1}^{n_m} \sum_{k=1}^n \left(x_k^{(i)} \right)^2$$

پالایش گروهی

۲۹

□ ایده. با داشتن ماتریس امتیازها و بردارهای $x^{(1)}, x^{(2)}, \dots, x^{(n_m)}$ می‌توان بردارهای $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n_u)}$ را تخمین زد

□ ایده. با داشتن ماتریس امتیازها و بردارهای $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n_u)}$ می‌توان بردارهای $x^{(1)}, x^{(2)}, \dots, x^{(n_m)}$ را تخمین زد

مقداردهی تصادفی



$\theta \rightarrow x \rightarrow \theta \rightarrow x \rightarrow \theta \rightarrow x \rightarrow \dots$

□ الگوریتم.

الگوريتم پالايس گروهي

پالایش گروهی

۳۱

ایده. تخمین $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n_u)}$ با داشتن بردارهای $x^{(1)}, x^{(2)}, \dots, x^{(n_m)}$ □

$$\min_{\theta^{(1)}, \dots, \theta^{(n_u)}} \frac{1}{2} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^n \left(\theta_k^{(j)} \right)^2$$

ایده. تخمین $x^{(1)}, x^{(2)}, \dots, x^{(n_m)}$ با داشتن بردارهای $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n_u)}$ □

$$\min_{x^{(1)}, \dots, x^{(n_m)}} \frac{1}{2} \sum_{i=1}^{n_m} \sum_{j:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{i=1}^{n_m} \sum_{k=1}^n \left(x_k^{(i)} \right)^2$$

هدف بهینه‌سازی در پالایش گروهی

۳۲

□ ایده. یادگیری **همزمان** بردارهای ویژگی $x^{(i)}$ و بردارهای $\theta^{(j)}$

□ تابع هدف.
 $J(x^{(1)}, \dots, x^{(n)}, \theta^{(1)}, \dots, \theta^{(n)}) =$

$$\frac{1}{2} \sum_{(i,j): r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{i=1}^{n_m} \sum_{k=1}^n \left(x_k^{(i)} \right)^2 + \frac{\lambda}{2} \sum_{i=1}^{n_u} \sum_{k=1}^n \left(\theta_k^{(j)} \right)^2$$

□ هدف.

$$\min_{\substack{x^{(1)}, \dots, x^{(n_m)} \\ \theta^{(1)}, \dots, \theta^{(n_u)}}} J(x^{(1)}, \dots, x^{(n)}, \theta^{(1)}, \dots, \theta^{(n)})$$

الگوریتم پالایش گرهی

۳۳

آموزش.

- مقداردهی اولیه به بردارهای x و θ با مقادیر تصادفی کوچک
- کمینه‌سازی تابع هزینه با استفاده از گرادیان کاهشی (یا روش‌های بهینه‌سازی پیشرفته)

$$x_k^{(i)} = x_k^{(i)} - \alpha \left(\sum_{j:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right) \theta_k^j + \lambda x_k^{(i)} \right)$$

$$\theta_k^{(j)} = \theta_k^{(j)} - \alpha \left(\sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right) x_k^{(i)} + \lambda \theta_k^{(j)} \right)$$

پیش‌بینی.

- برای کاربر j با بردار پارامتر $\theta^{(j)}$ و فیلم i با بردار ویژگی $x^{(i)}$

نرم‌السازی میانگین

کاربران جدید

۳۵

	Alice(1)	Bob(2)	Carol(3)	Dave(4)	eve(5)	
Titanic	5	5	0	0	?	0
Sound and Music	5	?	?	0	?	0
Casablanca	?	4	0	?	?	0
Fast and Furious	0	0	5	4	?	0
Desperado	0	0	5	?	?	0

$$\frac{1}{2} \sum_{(i,j): r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{i=1}^{n_m} \sum_{k=1}^n \left(x_k^{(i)} \right)^2 + \frac{\lambda}{2} \sum_{i=1}^{n_u} \sum_{k=1}^n \left(\theta_k^{(j)} \right)^2$$

$$\theta^{(5)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \rightarrow (\theta^{(5)})^T x^{(i)} = 0$$

نرمالسازی میانگین

۳۶

□ نرمالسازی میانگین.

$$Y = \begin{bmatrix} 5 & 5 & 0 & 0 & ? \\ 5 & ? & ? & 0 & ? \\ ? & 4 & 0 & ? & ? \\ 0 & 0 & 5 & 4 & ? \\ 0 & 0 & 5 & 0 & ? \end{bmatrix} \quad \mu = \begin{bmatrix} 2.5 \\ 2.5 \\ 2 \\ 2.25 \\ 1.25 \end{bmatrix} \rightarrow \quad Y_{norm} = \begin{bmatrix} 2.5 & 2.5 & -2.5 & -2.5 & ? \\ 2.5 & ? & ? & -2.5 & ? \\ ? & 2.0 & -2.0 & ? & ? \\ -2.25 & -2.25 & 2.75 & 1.75 & ? \\ -1.25 & -1.25 & 3.75 & -1.25 & ? \end{bmatrix}$$

□ پیش‌بینی: میزان علاقمندی کاربر j به فیلم i

$$\hat{y}(i, j) = (\theta^{(j)})^T x^{(i)} + \mu^{(i)}$$