



## فصل نهم : تشخیص هرزنامه

From: cheapsales@buystufffromme.com  
To: razavi@iust.ac.ir  
Subject: Buy now!

Deal of the week! Buy now!  
Rolex w4tchs - \$100  
Medicine (any kind) - \$50  
Also low cost M0rgages  
available.

From: Sullivan Kettler  
To: razavi@iust.ac.ir  
Subject: Christmas dates?

Hi Naser,  
Was talking to Nicolas about  
plans for Xmas. When do you get  
off work. Meet Dec 22?  
Sullivan.

## فصل نهم: ایجاد یک کلاس بند هرزنامه

□ یادگیری نظارت شده.

□  $x$ : ویژگی‌های ایمیل

■ انتخاب ۱۰۰ کلمه که نشان دهنده‌ی هرزنامه یا ایمیل معمولی هستند، مانند خرید، تخفیف، معامله، ...

□  $y$ : هرزنامه (۱) یا غیر هرزنامه (۰)

```
From: cheapsales@buystufffromme.com  
To: razavi@iust.ac.ir  
Subject: Buy now!
```

```
Deal of the week! Buy now!  
Rolex w4tchs - $100  
Medicine (any kind) - $50  
Also low cost M0rgages  
available.
```

□ **توجه.** در عمل معمولاً از  $n$  کلمه‌ی متداول‌تر (۱۰۰۰۰ تا ۵۰۰۰۰) در مجموعه‌ی آموزشی استفاده می‌شود.

## فصل نهم: ایجاد یک کلاس بند هرزنامه

□ س. چگونه یک کلاس بند با درصد خطای پایین ایجاد کنیم؟

□ جمع‌آوری داده‌های بسیار زیاد

□ توسعه‌ی ویژگی‌های پیشرفته بر اساس اطلاعات مسیریابی ایمیل

□ توسعه‌ی ویژگی‌های پیشرفته بر اساس کلمات به کار رفته در بدنه‌ی پیغام.

■ به عنوان مثال، آیا کلماتی مانند معامله و معاملات و معامله کننده باید به عنوان یک کلمه در نظر گرفته شوند یا خیر.

□ توسعه‌ی الگوریتم‌های پیچیده برای تشخیص غلط‌های املائی عمدی!



## فصل نهم : رویکرد پیشنهادی

□ انتخاب الگوریتم یادگیری و پیاده‌سازی.

□ با یک الگوریتم ساده که به سرعت قابل پیاده‌سازی باشد شروع کنید.

□ آن را پیاده‌سازی و بر روی مجموعه‌ی اعتبارسنجی آزمایش کنید.

□ عیب‌یابی الگوریتم یادگیری.

□ منحنی‌های یادگیری را ترسیم کنید تا بفهمید آیا نیاز به داده‌های بیشتر، ویژگی‌های بیشتر، کاهش ضریب تنظیم و غیره دارید یا خیر.

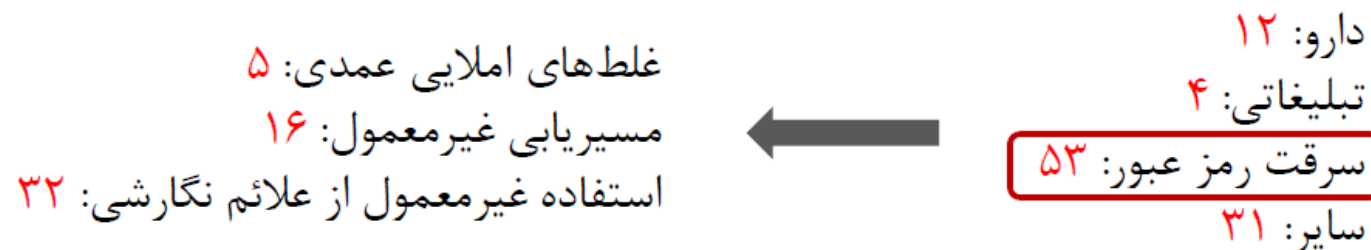
□ تحلیل خطا.

□ داده‌های موجود در مجموعه‌ی اعتبارسنجی را که الگوریتم در مورد آنها اشتباه کرده است، بررسی کنید.

□ ببینید آیا ویژگی مشترکی در این داده‌ها وجود دارد که باعث ایجاد خطا شده است یا خیر.

## فصل نهم: تحلیل خطا

- فرض کنید در مجموعه‌ی اعتبارسنجی ۵۰۰ نمونه ایمیل وجود دارد.
- الگوریتم یادگیری ۱۰۰ ایمیل را نادرست کلاس‌بندی کرده است.
- این ۱۰۰ نمونه را بررسی کنید و آنها را بر اساس مواردی مانند زیر دسته‌بندی کنید:
  - نوع ایمیل: دارو، تبلیغاتی، سرقت رمز عبور
  - سرنخ‌هایی که فکر می‌کنید می‌توانند به الگوریتم در کلاس‌بندی درست این ایمیل‌ها کمک کنند.



## فصل نهم: اهمیت ارزیابی های کمی و عددی

□ س. آیا کلماتی مانند **معامله**، **معاملات** و **معامله گر** باید یکسان در نظر گرفته شوند؟

□ برای این منظور می توان از نرم افزارهای مربوط به ریشه یابی کلمات استفاده نمود. [مانند Porter stemmer]

□ تحلیل خطا در این موارد کمک کننده نیست و تنها راه حل این است که ایده ی بالا را در عمل آزمایش کنیم.

□ به عبارت دیگر، نیاز داریم عملکرد الگوریتم را در هر دو مورد بر روی مجموعه ی اعتبارسنجی به صورت عددی ارزیابی کنیم و سپس بر اساس نتایج ارزیابی تصمیم گیری نماییم.

□ بدون ریشه یابی کلمات: ۵٪ خطا

□ با ریشه یابی کلمات: ۳٪ خطا



## سنجش خطا برای کلاس‌های نامتوازن

## فصل نهم: مثال (تشخیص سرطان)

□ یک مدل رگرسیون لجستیکی را آموزش دهید.

□ خروجی: سرطان ( $y = 1$ ); در غیر این صورت ( $y = 0$ )

□ فرض کنید خطای مدل آموزش داده شده برای مجموعه‌ی آزمایشی برابر با ۱٪ باشد.  
[۹۹٪ تشخیص درست]

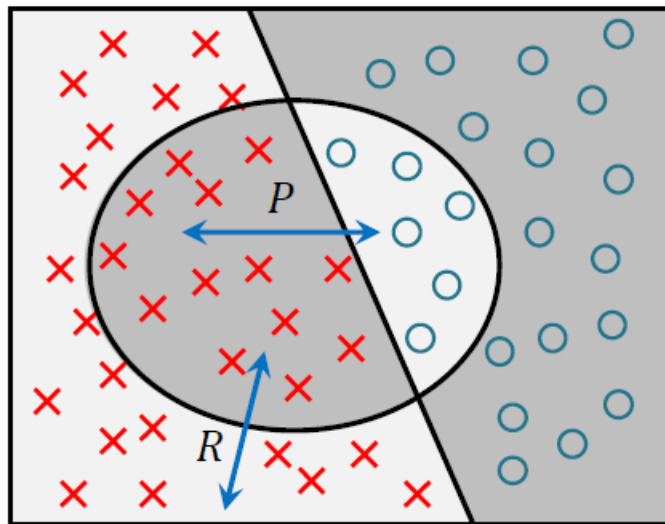
□ فرض کنید در داده‌های ما تنها ۰/۵ درصد از بیماران واقعاً سرطان دارند.

```
function y = predictCancer(x)
    y = 0; % just ignore x
end;
```

خطا: ۰/۵ درصد

کلاس نامتوازن. کلاسی که در آن نسبت تعداد نمونه‌های مثبت به تعداد نمونه‌های منفی (و یا بالعکس) بسیار کوچک (نزدیک به صفر) است.

## فصل نهم: نرخ درستی و نرخ یادآوری



## فصل نهم : نرخ درستی

□ **نرخ درستی.** نسبت تعداد نمونه‌هایی که به درستی مثبت تشخیص داده شده‌اند، به تعداد کل نمونه‌هایی که مثبت تشخیص داده شده‌اند.

		Actual Class	
		1	0
Predicted Class	1	True Positive	False Positive
	0	False Negative	True Negative

$$\frac{\text{True Positive}}{\# \text{ predicted Positive}} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

## فصل نهم : نرخ یادآوری

□ **نرخ یادآوری.** نسبت تعداد نمونه‌هایی که به درستی مثبت تشخیص داده شده‌اند، به تعداد کل نمونه‌هایی که واقعاً مثبت هستند.

		Actual Class	
		1	0
Predicted Class	1	True Positive	False Positive
	0	False Negative	True Negative

$$\frac{\text{True Positive}}{\# \text{ actual Positive}} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

## فصل نهم: موازنه میان نرخ درستی و نرخ یادآوری

$$\text{precision} = \frac{\text{true positives}}{\text{no. of predicted positive}}$$

$$\text{recall} = \frac{\text{true positives}}{\text{no. of actual positive}}$$

□ رگرسیون لجستیکی.

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

□ پیش‌بینی:

□  $y = 1$  اگر  $h_{\theta}(x) \geq 0.5$

□  $y = 0$  اگر  $h_{\theta}(x) < 0.5$

□ به منظور افزایش ضریب اطمینان می‌توان به صورت زیر پیش‌بینی نمود:

□  $y = 1$  اگر  $h_{\theta}(x) \geq 0.9$

□  $y = 0$  اگر  $h_{\theta}(x) < 0.9$

افزایش نرخ درستی

کاهش نرخ یادآوری



## فصل نهم: موازنه میان نرخ درستی و نرخ یادآوری

$$\text{precision} = \frac{\text{true positives}}{\text{no. of predicted positive}}$$

$$\text{recall} = \frac{\text{true positives}}{\text{no. of actual positive}}$$

□ رگرسیون لجستیکی.

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

□ پیش‌بینی:

□ اگر  $y = 1$ :  $h_{\theta}(x) \geq 0.5$

□ اگر  $y = 0$ :  $h_{\theta}(x) < 0.5$

□ اگر بخواهیم تعداد نمونه‌های مثبت کمتری را از دست بدهیم:

□ اگر  $y = 1$ :  $h_{\theta}(x) \geq 0.3$

□ اگر  $y = 0$ :  $h_{\theta}(x) < 0.3$



کاهش نرخ درستی

افزایش نرخ یادآوری

## فصل نهم : امتیاز F

□ س. چگونه می‌توان نرخ درستی و نرخ یادآوری الگوریتم‌های مختلف را با هم مقایسه نمود؟

نرخ یادآوری	نرخ درستی	
۰/۴	۰/۵	الگوریتم ۱
۰/۱	۰/۷	الگوریتم ۲
۱/۰	۰/۰۲	الگوریتم ۳

□ امتیاز  $F$ .

□ اگر  $P = 0$  یا  $R = 0$ ، آنگاه امتیاز  $F$  برابر با صفر است.

□ اگر  $P = 1$  و  $R = 1$ ، آنگاه امتیاز  $F$  برابر با یک است.

$$2 \frac{P \cdot R}{P + R}$$



## فصل نهم : امتیاز F

□ س. چگونه می‌توان نرخ درستی و نرخ یادآوری الگوریتم‌های مختلف را با هم مقایسه نمود؟

امتیاز F	نرخ یادآوری	نرخ درستی	
۰/۴۴۴	۰/۴	۰/۵	الگوریتم ۱
۰/۱۷۵	۰/۱	۰/۷	الگوریتم ۲
۰/۰۳۹	۱/۰	۰/۰۲	الگوریتم ۳

□ امتیاز  $F$ .

□ اگر  $P = 0$  یا  $R = 0$ ، آنگاه امتیاز  $F$  برابر با صفر است.

□ اگر  $P = 1$  و  $R = 1$ ، آنگاه امتیاز  $F$  برابر با یک است.

$$2 \frac{P \cdot R}{P + R}$$

## داده‌ها برای یادگیری ماشین

۱۸

## فصل نهم : طراحی یک سیستم یادگیری با دقت بالا

❑ مسئله. تشخیص کلمات مشابه [بانکو و بریل، ۲۰۰۱]

{to, too, two}, {then, than}

For breakfast, I ate \_\_\_\_\_ eggs.

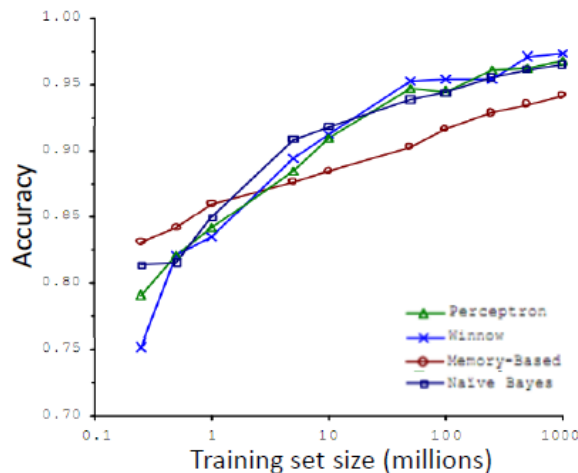
❑ الگوریتم‌ها.

❑ پرسپترون (رگرسیون لجستیکی)

❑ وینو

❑ مبتنی بر حافظه

❑ کلاس‌بند بیز



« برنده آن کسی نیست که بهترین الگوریتم را در اختیار دارد. برنده آن کسی است که داده‌های بیشتری دارد. »

## فصل نهم : منطق توجیه کننده برای داده های زیاد

□ توجه. داشتن داده‌های بیشتر تنها زمانی مفید است که بردار ویژگی  $x$  دربرگیرنده‌ی اطلاعات کافی برای تخمین خروجی  $y$  باشد.

□ مثال ۱. پر کردن جای خالی با کلمات داده شده (بله)

□ مثال ۲. تخمین قیمت خانه تنها با دانستن اندازه‌ی آن (خیر)

□ یک آزمایش مفید. آیا با داشتن ورودی  $x$ ، یک انسان خبره می‌تواند مقدار  $y$  را با ضریب اطمینان بالا پیش‌بینی کند؟

## فصل نهم : منطق توجیه کننده برای داده های زیاد

□ استفاده از یک الگوریتم یادگیری قدرتمند با پارامترهای زیاد.

□ رگرسیون خطی یا لجستیکی با تعداد بسیار زیادی از ویژگی ها

□ شبکه‌ی عصبی با تعداد بسیار زیادی از واحدهای مخفی

□ خطای مجموعه‌ی آموزشی کم (۱)

□ استفاده از یک مجموعه آموزشی بسیار بزرگ. [کاهش خطر بیش‌برازش]

□ خطای مجموعه آموزشی تقریباً برابر با خطای مجموعه‌ی آزمایشی (۲)

□ نتیجه (۱) و (۲). خطای مجموعه‌ی آزمایشی کم. [قابلیت تعمیم بالا]

باتشکر از توجه شما

