یادگیری ماشین

تمرین «۳»

دستیاران آموزشی

آیدین کیانی

مرضیه باقری‌نیا

مهدی حسینی چاگاهی

دکتر سامان هراتی‌زاده

دانشگاه تهران - دانشکده سامانه‌های هوشمند

نیم‌سال اول ۱۴۰۳-۱۴۰۲

ددلاین: ساعت ۲۳:۵۹ | ۱۴۰۲/۰۹/۲۷

راهنمای تحویل

قبل از پاسخ دادن به پرسش‌ها، موارد زیر را با دقت مطالعه نمایید:

- از پاسخ‌های خود یک گزارش در قالبی که در صفحه‌ی درس در سامانه‌ی Elearn با نام **REPORTS_TEMPLATE.docx** قرار داده شده تهیه نمایید.
- **کیفیت گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است**؛ بنابراین، لطفا تمامی نکات و فرض‌هایی را که در پیاده‌سازی‌ها و محاسبات خود در نظر می‌گیرید در گزارش ذکر کنید.
- در گزارش خود مطابق با آنچه در قالب نمونه قرار داده شده، برای شکل‌ها زیرنویس و برای جدول‌ها بالانویس در نظر بگیرید.
- الزامی به ارائه توضیح جزئیات کد در گزارش نیست، اما باید نتایج بدست آمده از آن را گزارش و تحلیل کنید.
- **تحلیل نتایج الزامی می‌باشد، حتی اگر در صورت پرسش اشاره‌ای به آن نشده باشد.**
- **کدهای ارسالی می‌بایست قابلیت اجرای دوباره داشته باشند، با این حال، دستیاران آموزشی ملزم به اجرای کدهای شما نیستند**؛ بنابراین، هرگونه نتیجه و یا تحلیلی که در صورت پرسش از شما خواسته شده را به طور واضح و کامل در گزارش بیاورید. در صورت عدم رعایت این مورد، بدیهی است که از نمره تمرین کسر می‌شود.
- **در صورت استفاده از Jupyter لازم است تا تمامی کد اجرا شود و خروجی هر سلول حتما در این فایل ارسالی شما ذخیره شده باشد در غیر اینصورت ورودی‌ها و خروجی‌ها متناظر می‌بایست در گزارش آورده شوند.** بنابراین برای مثال اگر خروجی سلولی یک نمودار است که در گزارش آورده‌اید، این نمودار باید هم در گزارش هم در نوت‌بوک کدها وجود داشته باشد.
- **با این که بحث در مورد تمرین‌ها منعی ندارد اما راه‌حل شما می‌بایست توسط شما (و فقط شما) باشد. همچنین، تمامی مطالب جانبی در گزارش باید رفرنس داده شود. یادآوری می‌شود که عدم صداقت علمی[1] عواقب شدیدی را به همراه دارد.**
- **استفاده از کدهای آماده برای تمرین‌ها به هیچ وجه مجاز نیست.**
- **در صورت مشاهده‌ی تقلب امتیاز تمامی افراد شرکت‌کننده در آن، به میزان بارم سوال نمره منفی لحاظ می‌شود.**
- لطفا گزارش، کدها و سایر ضمایم را به در یک پوشه با نام زیر قرار داده و آن را فشرده سازید، سپس در سامانه‌ی Elearn بارگذاری نمایید:

HW[Number]_[Lastname]_[StudentNumber].zip

در صورت وجود سوال، ابهام و یا درخواست راهنمایی با دستیاران آموزشی مرتبط با هر پرسش از طریق ایمیل‌های آورده شده در سربرگ در ارتباط باشید.

[1] Academic dishonesty

یادگیری ماشین

تمرین «۳»

دستیاران آموزشی

آیدین کیانی

مرضیه باقری‌نیا

مهدی حسینی چاگاهی

دکتر سامان هراتی‌زاده

دانشگاه تهران - دانشکده سامانه‌های هوشمند

نیم‌سال اول ۱۴۰۳-۱۴۰۲

ددلاین: ساعت ۲۳:۵۹ | ۱۴۰۲/۰۹/۲۷

فهرست سوالات

شکل‌ها

یادگیری ماشین

تمرین «۳»

دستیاران آموزشی

آیدین کیانی

مرضیه باقری‌نیا

مهدی حسینی چاگاهی

دکتر سامان هراتی‌زاده

دانشگاه تهران - دانشکده سامانه‌های هوشمند

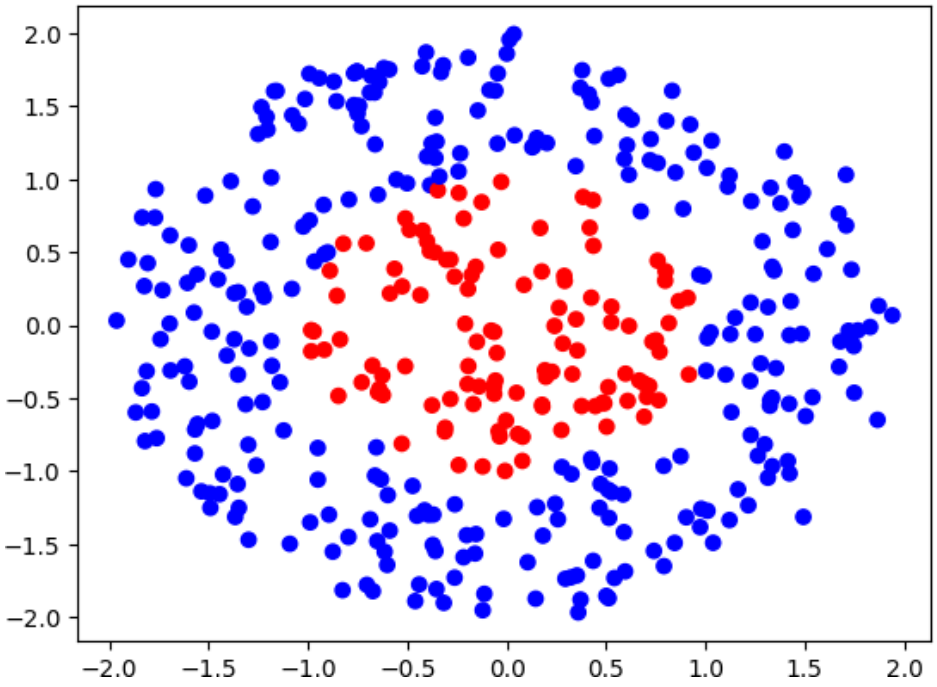
نیم‌سال اول ۱۴۰۳-۱۴۰۲

ددلاین: ساعت ۲۳:۵۹ | ۱۴۰۲/۰۹/۲۷

جداپذیری خطی

در این تمرین میخواهیم با مفهوم جداپذیری خطی و تغییر فضای ویژگی آشنا شویم.

1. ابتدا یک مجموعه داده با 2 ویژگی x و y و به همان تعداد برچسب مانند شکل زیر تولید و نمایش دهید.



2. حال سعی کنید با استفاده از Logistic Regression آن را جدا کنید. مرز تصمیم را رسم و دقت را گزارش کنید.

3. در این مرحله مجموعه‌داده را به یک فضای جدید انتقال دهید پس از نمایش در فضای جدید، بخش ۲ را تکرار کنید (راهنمایی: میتوانید در مورد دستگاه مختصات قطبی تحقیق کنید).

در ادامه برای یک مسئله تک ویژگی gradient descent را پیاده‌سازی کنید. برای این منظور دو فایل gradient.py و gradient-descent.ipynb را تکمیل کرده و روابط پیدا کردن theta_0 و theta_1 را بنویسید (در صورت وجود دو روش پیاده‌سازی «با حلقه» و «بدون حلقه» نمره‌ی کامل به شما تعلق می‌گیرد).

پیش‌بینی قیمت خودرو با استفاده از رگرسیون خطی

مجموعه دادگان داده شده مربوط به تعدادی خودرو است که ویژگی‌های آن به شکل زیر است:

Make: The company of the car.
Model: The name of the car.
Price: Selling price of the car in INR.

Year: Manufacturing year of the car.
Kilometes: Total kilometers driven.
Fuel Type: The fuel type of the car.

Transmission: Gear transmission of the car.
Location: City in which the car is being sold.

یادگیری ماشین

تمرین «۳»

دستیاران آموزشی

آیدین کیانی

مرضیه باقری‌نیا

مهدی حسینی چاگاهی

دکتر سامان هراتی‌زاده

دانشگاه تهران - دانشکده سامانه‌های هوشمند

نیم‌سال اول ۱۴۰۳-۱۴۰۲

ددلاین: ساعت ۲۳:۵۹ | ۱۴۰۲/۰۹/۲۷

Color: Color of the car.
Owner: Number of previous owners.
Seller Type: Indicates if the car is sold by an individual or a dealer.
Engine: Engine capacity of the car in cc.

Max Power: Maximum power in bhp@rpm.
Max Torque: Maximum torque in Nm@rpm.
Drivetrain: AWD/RWD/FWD.
Length: Length of the car in mm.
Width: Width of the car in mm.

Height: Height of the car in mm.
Seating Capacity: Maximum number of people that can fit in a car.
Fuel Tank Capacity: Maximum fuel capacity of the car in litres.

۱. **ایجاد داده‌های آزمون**

قبل از هر چیز، داده‌های آزمون خود را جدا کنید. (داده‌ها را به گروه‌های درآمدی تقسیم کنید و از `stratify` در تابع `train_test_split` برای ایجاد مجموعه آزمون نمونه‌ای استفاده کنید).

۲. **کاوش در دادگان**

تحلیل داده‌ها را انجام دهید تا به دانشی دست یابید. به عنوان مثال، برای ویژگی‌های مختلف هیستوگرام بسازید و میانگین قیمت مدل‌های مختلف را محاسبه کنید. از تحلیل‌های جذاب دیگری نیز استفاده کنید.

۳. **مقدارهای خالی**

درصد خانه‌های خالی هر ستون را مشخص کنید و راه‌هایی برای پر کردن آنها پیشنهاد دهید. دلیل انتخاب خود را بیان کنید. به طور خلاصه درباره انواع روش‌های پر کردن خانه‌های خالی در داده‌های عددی و غیر عددی توضیح دهید. مطالعه [منبع] مناسب است.

۴. **ماتریس همبستگی**

ماتریس همبستگی داده را رسم کنید. نتایج را توضیح دهید و بیان کنید چرا برخی از ویژگی‌ها همبسته هستند و این همبستگی چه معنایی دارد. به یاد داشته باشید که مدل‌های خطی، به ویژگی‌های با همبستگی قوی خوب پاسخ نمی‌دهند. سعی کنید ویژگی‌های جدیدی از ویژگی‌های همبسته تولید کنید.

۵. **تبدیل به اعداد**

داده‌های غیر عددی را با یک روش مناسب تبدیل کنید. تلاش کنید از تابع `get_dummies` استفاده نکنید تا مشکلات احتمالی در زمان آزمون را جلوگیری کنید. برای مطالعه بیشتر به [منبع مناسب] مراجعه کنید.

۶. **تنظیم پارامترها**

با استفاده از داده‌های آموزش و `GridSearchCV` بهترین پارامترها را برای مدل‌های خطی عادی مانند Lasso و Ridge به دست آورید. به یاد داشته باشید که تا قبل از مرحله ارزیابی با داده‌های آزمون کار نکنید و بهترین مدل را انتخاب کنید.

۷. **سنجه‌های ارزیابی**

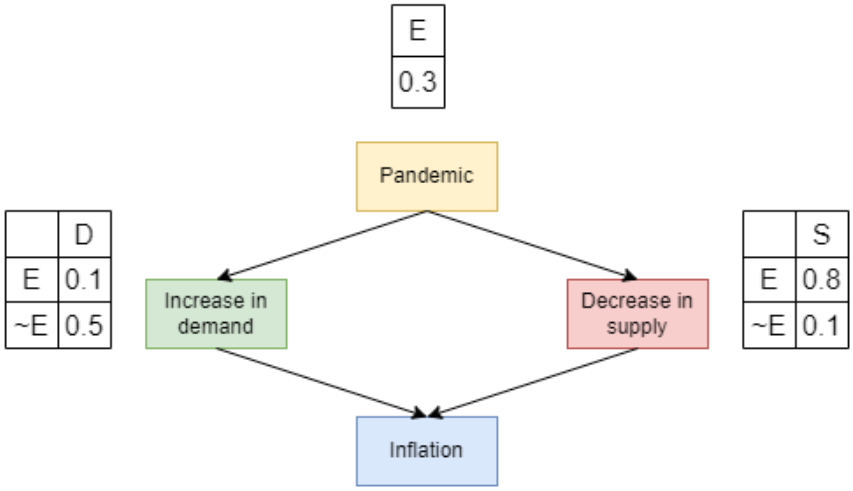
در مورد RMSE و امتیاز R2 تحقیق کنید و مقادیر آنها را برای داده‌های آزمون محاسبه کنید. [لینک مناسب] را به عنوان مرجع مطالعه استفاده کنید.

۸. **ارزیابی متقابل k-fold**

به صورت خلاصه در مورد دلیل استفاده از ارزیابی متقابل k-fold توضیح دهید. برای مدل خود با k=5 مقادیر آن را محاسبه کرده و Boxplot آن را رسم کنید.

شبکه‌های Bayesian

در زمان شیوع پاندمی کرونا، افزایش نقدینگی افراد جامعه و در نتیجه آن افزایش تقاضا، به موازاتِ توقف یا کاهش تولید کسب و کار ها که ناشی از سیاست‌های قرنطینه بود، باعث ایجاد تورم، به ویژه در کالاها، در سراسر جهان شد و نتایج آن اقتصاد جهان را رو به بحران برد. در نتیجهِ این تحلیل، تصویر زیر حاوی شبکه روابط میان عوامل ایجاد تورم و نیز جدول احتمال شرطی[1] مربوط به هر عامل است؛ با توجه به این شبکه، به سوالات زیر پاسخ دهید:



شکل ۱

الف) احتمال ایجاد تورم پس از ایجاد یک بحران یا پاندمی را محاسبه نمایید.

ب) احتمال اینکه رخدادِ تورم ناشی از ایجاد یک بحران یا پاندمی باشد، چقدر است؟

[1]. Conditional Probability Table

یادگیری ماشین

تمرین «۳»

دستیاران آموزشی

آیدین کیانی

مرضیه باقری‌نیا

مهدی حسینی چاگاهی

دکتر سامان هراتی‌زاده

دانشگاه تهران - دانشکده سامانه‌های هوشمند

نیم‌سال اول ۱۴۰۳-۱۴۰۲

ددلاین: ساعت ۲۳:۵۹ | ۱۴۰۲/۰۹/۲۷

پیش‌بینی ریزش مشتری

ریزش مشتری به زمانی اشاره دارد که مشتری به رابطه خود با یک تجارت[1] پایان می‌دهد. از آنجایی که جذب مشتریان جدید می‌تواند چندین برابر پرهزینه‌تر از نگهداری مشتریانِ فعلی باشد، درک اینکه چه چیزی باعث ریزش مشتریان می‌شود و همچنین توانایی شناسایی دقیق مشتریانی که در معرض خطر بالای ریزش هستند، بسیار مهم است و ممکن است به کسب‌وکارها در ایجاد استراتژی‌های بازاریابی مناسب و حفظ مشتریان خود کمک کند. در این سوال قصد داریم تا مساله پیش‌بینی ریزش مشتریان یک شرکت مخابراتی را حل نماییم؛ بنابراین لازم است تا جهت پاسخ به این سوال مراحل زیر را گام به گام پیاده‌سازی نمایید:

- فایل customer_churn.csv حاوی اطلاعات بیش از ۷۰۰۰ نمونه به همراه ۲۰ ویژگی می‌باشد. در گام اول لازم است تا داده‌های درون این فایل را دریافت و ذخیره نمایید.
- پس از ذخیره داده‌ها، لازم است تا پیش‌پردازش لازم را بر روی آن‌ها انجام دهید و داده‌های تمیز را به مدل وارد نمایید؛ دقت کنید که پیش‌پردازش شما باید شامل موارد زیر باشد:
  - رسیدگی به داده‌های ناقص و یا از دست‌رفته
  - رسیدگی به یکپارچگی و درستی نوع داده‌ها؛ برای مثال ممکن است داده‌ای ماهیت عددی داشته باشد اما به صورت رشته‌ای ذخیره شده باشد.
  - تبدیل داده‌های دسته‌ای[2] به داده‌های عددی
  - استفاده از ابزارهای تحلیلی و نمایشی به منظور بررسی تاثیر ویژگی‌ها بر هم و بر متغیر هدف؛ بدین منظور می‌توانید از متدهایی مانند pairplot استفاده نمایید.
- پس از پردازش داده‌ها، توزیع داده‌ها بر حسب متغیر هدف را رسم نمایید؛ آیا داده‌ها نسبت به متغیر هدف متعادل[3] هستند؟ توضیح دهید که داده‌های نامتعادل چه تاثیری بر عملکرد مدل می‌گذارند و از چه روش‌هایی برای رفع مشکل داده‌های نامتعادل[4] استفاده می‌شود؟ همچنین در صورتی که داده‌های شما متعادل نیست، یک روش را به انتخاب خود برای متعادل‌سازی انتخاب نموده و داده‌ها را متعادل نمایید.
- پس از آماده شدن داده‌ها، از یک شبکه بیزی برای آموزش مدل پیش‌بین بر روی داده‌ها استفاده نمایید؛ برای پیاده‌سازی شبکه بیزی می‌توانید از کتابخانه pgmpy استفاده نمایید (اما اجباری به استفاده از آن وجود ندارد).

---

[1]. Business
[2]. Categorical
[3]. Balanced
[4]. Unbalanced

یادگیری ماشین

تمرین «۳»

دستیاران آموزشی

آیدین کیانی

مرضیه باقری‌نیا

مهدی حسینی چاگاهی

دکتر سامان هراتی‌زاده

دانشگاه تهران - دانشکده سامانه‌های هوشمند

نیم‌سال اول ۱۴۰۳-۱۴۰۲

ددلاین: ساعت ۲۳:۵۹ | ۱۴۰۲/۰۹/۲۷

- در انتها نیز، مدل پیش‌بین آموزش دیده را بر روی داده‌های تستِ از پیش آماده شده، بیازمایید و نتیجه را با استفاده از ابزارها و معیارهای مناسب گزارش دهید.

حقه کرنل[1]

۱. نشان دهید که برای کرنل $K(x_i,x_j) = \exp(-1/2\ ||x_i - x_j||^2)$ به ازای هر دو ورودی دلخواه در فضای feature space خواهیم داشت:

$$||\varphi(x_i) - \varphi(x_j)|| \leqslant 2$$

۲. اگر کرنل های معتبر $k_1(x,y)$, $k_2(x,y)$ را داشته باشیم، اعتبار کرنل های زیر را به کمک تئوری mercer بررسی نمایید.

$$1)\ K(x,y) = f(x)K_1(x,y)f(y),\ \ for\ \ \forall f$$
$$2)\ K(x,y) = X^TAy,\ \ A\ is\ positive\ definite.$$
$$3)\ K(x,y) = (x^Ty+1)^p,\ p>0$$
$$4)\ K(x,y) = exp(K_1(x,y))$$

معادله تابع تفکیک[2]

مجموعه داده مطابق جدول ۱ را در نظر بگیرید.

جدول ۱: نمونه داده

x	Label
0	+
1	-
-1	-

۱. آیا دو کلاس مشخص شده به صورت خطی جداپذیرند؟

۲. فرض کنید هر نقطه از این فضا را به یک نقطه در فضای سه بعدی با تابع φ ، که در زیر تعریف شده است، نگاشت کنیم. آیا کلاس ها در این حالت جداپذیر خطی هستند؟ در صورت جداپذیر بودن صفحه جداکننده را پیدا کنید.

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

۱. یک متغیر برای کلاس ها به صورت $y_i \in \{-1, 1\}$ در نظر بگیرید که کلاس هر کدام از x_i ها را نشان می دهد و داریم $W = (W_1, W_2, W_3)^T$ ، با طبقه بند max-margin SVM مسئله بهینه سازی زیر را حل کنید.

$$\underset{w,b}{\text{Min}}\ \frac{1}{2}||W||^2$$

$$s.t.\quad y_i(W^T\varphi(x_i) + b \geq 1,\quad i = 1,2,3$$

با استفاده از روش ضرایب لاگرانژ W, b و اندازه margin را بیابید.

ماشین بردار پشتیبان

۱. در این سوال به اعمال طبقه بند SVM بر روی مجموعه داده ی Q.zip می‌پردازیم.

[1] Kernel Trick
[2] Discriminant Function Equation

یادگیری ماشین

تمرین «۳»

دستیاران آموزشی

آیدین کیانی

مرضیه باقری‌نیا

مهدی حسینی چاگاهی

دکتر سامان هراتی‌زاده

دانشگاه تهران - دانشکده سامانه‌های هوشمند

نیم‌سال اول ۱۴۰۳-۱۴۰۲

ددلاین: ساعت ۲۳:۵۹ | ۱۴۰۲/۰۹/۲۷

۲. در مورد کرنل های linear, RBF, sigmoid, and polynomial تحقیق کنید و بیان کنید هر کدام برای طبقه بندی چه مجموعه داده هایی مناسب هستند.

۳. در این مرحله با هر کدام از روشهای زیر و ویژگی های داده شده، طبقه بندی را انجام دهید و برای هرکدام دقت طبقه بندی داده های آموزش و تست را تا چهار رقم اعشار در قالب یک جدول گزارش نمایید.

SVM with RBF Kernel, C = 1, 100, 1000
SVM with Linear Kernel, C = 1, 100, 1000
SVM with Polynomial Kernel, C = 1, 100
SVM with Sigmoid Kernel, C = 1, 100

٤. در این مرحله قصد داریم که مقادیر بهینه هایپر پارامترها را بیابیم. برای این کار از Grid Search استفاده می کنیم. سعی کنید که از بین مقادیر پارامترهایی که در زیرمشخص شده اند، با بهره گیری از روش Grid Search بهترین پارامترها را گزارش کنید.(گزارش پارامترها شامل نوع کرنل و مقادیر [C, gamma] می باشد).

Kernel: RBF, C = [1, 10, 100, 500], gamma = [0.1, 0.3, 0.5, 0.7, 0.9]
Kernel: Linear, C = [1, 10, 100, 1000]
Kernel: Polynomial, degree = [2, 3, 4], C = [1, 10, 100, 500], gamma = [0.01, 0.03, 0.05]

۵. بهترین طبقه بند بخش ب و پ را بر روی داده ها اعمال کنید و مقادیر accuracy, recall, precision, and f1-score را گزارش نمایید.

رگرسیون با ماشین بردار پشتیبان[1]

در این سوال روش SVR را بررسی خواهیم کرد.

۱. در این سوال به کمک مجموعه داده میزان درآمد افراد بر اساس موقعیت شغلی آن ها، می خواهیم یک مسئله Support Vector Regression را حل کنیم. این مجموعه داه شامل سه ستون است که در ستون اول موقعیت شغلی آن ها شرح داده شده، در ستون دوم رتبه کاری و در ستون سوم میزان درامد آورده شده است. در این سوال به کمک سه کرنل RBF, linear, and polynomial میزان درآمد را پیش بینی نمایید و در خروجی مقادیر تخمین زده شده و واقعی را در یک نمودار نمایش دهید.

۲. در این بخش می خواهیم مسئله SVR را حل کنیم که هدف آن پیش بینی هزینه اقامت در هتل به ازای ویژگی های مختلف می باشد. مجموعه داده در فایل Hotel.zip قرار داده شده است. از داده های فایل H1.csv به عنولن داده های آموزش و از داده های فایل H2.csv به عنوان داده های تست استفاده نمایید. در این سوال از تمامی ویژگی های numeric و categorical استفاده نمایید و هزینه اقامت در هتل را پیش بینی کنید و در یک فایل csv ذخیره نمایید. این فایل باید شامل سه ستون مقدار واقعی، مقدار پیش بینی شده و اختلاف بین مقدار واقعی با مقدار پیش بینی شده باشد.

تذکر۱: پیش پردازش های لازم بر روی دیتاست انجام شود.

تذکر۲: در این سوال هدف پیش بینی Average Daily Rate (ADR) است.

[1] SVR