هوش مصنوعی **تمرین شماره‌ی 4**

مدرسین: **دکتر فدایی و دکتر یعقوب‌زاده**

طراحان: پرهام سازدار، اولدوز نیساری، محمد امانلو

مهلت تحویل: 19 اردیبهشت ماه ۱۴۰۳، ساعت ۲۳:۵۹

بخش کتبی

مبحث اول

سوال اول

نادقیقی، تمیز نبودن، نقص داشتن و وجود داده‌های خارج از محدوده و نامعقول از جمله مشکلاتی هستند که در داده‌هایی که از دنیای واقعی به دست می‌آوریم وجود دارند. به نظر شما برای حل مشکلات زیر چه راهکارهایی ارائه می‌شوند؟

1) وجود نداشتن یک یا چند ویژگی در داده‌های آموزش
2) نامتعادل بودن توزیع داده‌ها در کلاس‌ها
3) وجود نویز در داده‌ها
4) وجود ویژگی‌های وابسته

پاسخ :

1) وجود نداشتن یک یا چند ویژگی در داده‌های آموزش:

الف) برای حل مشکل نقص داشتن، اگر هنوز در مرحله جمع‌آوری داده‌ها هستیم، یک روش افزایش تعداد داده‌ها با استفاده از منابع دیگر است.

ب) اگر از مرحله جمع‌آوری داده عبور کرده‌ایم، باز هم با روش‌های مبتنی بر هوش مصنوعی می‌توانیم داده‌های متناسب جدیدی ساخت و از آن‌ها بهره برد.

ج) برای جبران داده‌های حذف شده، می‌توانیم ستون‌ها و ویژگی‌های جدید بر مبنای ستون‌های دیگر بسازیم تا در آن ستون جدید بتوانیم داده‌های حذف شده را برحسب ستون‌های دیگر پر کنیم.

د) اگر ستونی وجود داشته باشد که داده‌های حذف شده زیادی داشته باشد، می‌توانیم از کل ستون صرف نظر کنیم و آن را حذف کنیم.

ه) داده‌های از دست رفته و ناموجود را می‌توانیم با انواع روش‌ها (از روش‌های پیش‌بینی‌کننده تا پر کردن با میانگین و میانه) پر کنیم.

2) نامتعادل بودن توزیع داده‌ها در کلاس‌ها
الف) یک روش دیگر استفاده از انواع روش‌های نمونه‌گیری مجدد است. با افزایش تعداد داده‌ها از اثر نامتعادل بودن کاسته می‌شود.

ب) روش دوم می‌تواند تغییر وزن و اثر بخشی ویژگی‌ها باشد که با الگوریتم‌هایی نظیر svm می‌توانیم این کار را انجام دهیم (البته در زمان model کردن هم می‌توانیم از ورژن تغییر یافته الگوریتم‌ها مثل balanced random forest استفاده کنیم).

3) وجود نویز در داده‌ها :

الف) برای شناسایی این نویزها شناسایی outlierها بسیار مفید است و به این منظور می‌توان از روش های آماری مثل z-score استفاده کرد.

ب) دسته‌ای از روش‌ها وجود دارد که اصطلاحا به آن‌ها روش‌های smoothing می‌گویند. یکی از مشهورترین این روش‌ها Gaussian smoothing است که بر اساس توزیع گوسی یک میانگین‌گیری وزن‌دار انجام می‌دهد.

ج) مشابه حالت قبل می‌توانیم از مدل‌هایی که تاثیرپذیری کمتری نسبت به نویز دارند (مثل random forest, gradient boosting) استفاده کنیم.

4) وجود ویژگی‌های وابسته :
الف) یک روش ساده می‌تواند ترکیب کردن ویژگی‌های وابسته و ساختن یک ویژگی جدید باشد.
ب) استفاده از ابزارهای تحلیل و مناسب‌سازی داده‌ها یک روش دیگر است. مثلا استفاده از Principal Component Analysis که بعد داده‌ها را کم می‌کند.
ج) استفاده از روش‌هایی که وزن کمتری به ویژگی‌های وابسته می‌دهند و نقش آن‌ها را کمرنگ‌تر می‌کنند مثل روش‌های regularization.

سوال دوم

یک مشاور تحصیلی در حال بررسی روی یک مجموعه داده درباره ساعت مطالعاتی دانشجویان و نمرات آزمون‌هایشان است. او توانسته است معادله رگرسیون خطی زیر را با توجه به داده‌های موجود به دست آورد:
نمره آزمون = 60 + 5 * ساعت مطالعاتی
اما با توجه به تاثیر انکار ناپذیر آزمون دادن در آمادگی دانشجویان او قصد دارد که نقش این مسئله را هم در نمره آزمون در نظر بگیرد . به نظر شما او چه مدل ریاضیاتی برای درک ارتباط بین این دو ویژگی پیشنهاد خواهد داد؟ با استفاده از least square method سعی کنید توضیح دهید چگونه ضرایب مناسب را پیدا می کند؟ اگر از gradient descent استفاده کند چطور؟ آیا تکنیک دیگری برای کم کردن اختلاف مجموع مربعات و مقادیر مشاهده شده می‌شناسید؟

پاسخ:

مدل ریاضیاتی که هم‌اکنون استفاده می‌کنید رگرسیون خطی است. با اضافه شدن مولفه تعداد آزمون‌ها برای دانشجو مدل ریاضیاتی به صورت زیر تغییر می‌کند:

1) تبدیل شدن به multiple regression:

نمره آزمون = ساعت مطالعاتی * 5+ c * تعداد آزمون ها + c1

2) با در نظر گرفتن بیشتر تاثیر هر یک می توانیم به صورت زیر عمل کنیم :

نمره آزمون = ساعت مطالعاتی * 5+ c * تعداد آزمون ها * تعداد آزمون ها + c2

یا:

نمره آزمون = ساعت مطالعاتی * ساعت مطالعاتی * 5+ c * تعداد آزمون ها + c3

حال برای به دست آوردن ضرایب به کمک least square method از روش زیر استفاده می‌کنیم: (این روش با تحقیق درباره روش های محاسباتی به دست آمده است ، هر چند توضیح ساده فرمول‌های least square method هم برای این روش کافی است.)

1) تعریف ماتریسی به صورت زیر:

$$X = \begin{bmatrix} 1 & \text{Study hours}_1 & \text{ExamNum}_1 \\ 1 & \text{Study hours}_2 & \text{ExamNum}_2 \\ \vdots & \vdots & \vdots \\ 1 & \text{Study hours}_n & \text{ExamNum}_n \end{bmatrix}$$

تعریف ماتریس نتیجه به صورت زیر:

$$y = \begin{bmatrix} \text{Exam score}_1 \\ \text{Exam score}_2 \\ \vdots \\ \text{Exam score}_n \end{bmatrix}$$

حال ماتریس ضرایب طبق فرمول زیر به دست می‌آید:

$$(X^T X)^{-1} X^T y$$

حال روشی را بررسی می‌کنیم که به کمک gradient decent بخواهیم ضرایب را به دست آوریم.

به این منظور در ابتدا ضرایبی را در ابتدا در نظر می‌گیریم. سپس تابع هزینه را تعریف می‌کنیم:

$$J(\beta_0, \beta_1, \beta_2) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2$$

سپس ضرایب را مطابق فرمول‌های زیر به دست می آوریم:

$$c_0 := c_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i)$$

$$c_1 := c_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i) \cdot \text{Study hours}_i$$

$$c_2 := c_2 - \alpha \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i) \cdot \text{ExamNum}_i$$

در این مرحله با تعیین ضریب آلفا تا رسیدن به نتیجه مطلوب این مراحل را تکرار می‌کنیم.

تکنیک‌های دیگر برای کم کردن اختلاف مجموع مربعات و مقادیر مشاهده شده:

1) SGD
2) bayesian reqgression: این تکنیک با در نظر گرفتن تاثیر باور قبلی دقت مدل را افزایش می‌دهد و همین باعث کم کردن اختلاف مجموع مربعات و مشاهدات می‌شود.
3) Lasso Regression: یک روش regularization است که تاثیر overfitting را کم کرده و منجر به دقت بیشتر می‌شود.

سوال سوم

ارزیابی مدلی که برای پیش‌بینی استفاده کرده‌ایم بسیار ضروری است. فرض کنید برای پیش‌بینی spam بودن از مدل بر مبنای logistic regression زیر استفاده کرده‌ایم:

$$p(Spam) = \frac{1}{1+e^{-(ß_0+ß_1 * email\ Length)}}$$

و فرض کنید نتایج به صورت

TP : 300 , TN :200 , FP :30 , FN : 20

باشد.
برای ارزیابی این مدل ابتدا confusion matrix را رسم کنید، سپس precision, recall، accuracy و F1-score را به دست آورید.

پاسخ :

	Predicted Spam (Positive)	Predicted Not Spam (Negative)
Actual Spam (Positive)	True Positive (TP): 300	False Negative (FN): 20
Actual Not Spam (Negative)	False Positive (FP): 30	True Negative (TN): 200

Recall = $\frac{TP}{TP+FN} = \frac{300}{300+20} = 0.9375$

Precision = $\frac{TP}{TP+FP} = \frac{300}{300+30} = 0.9091$

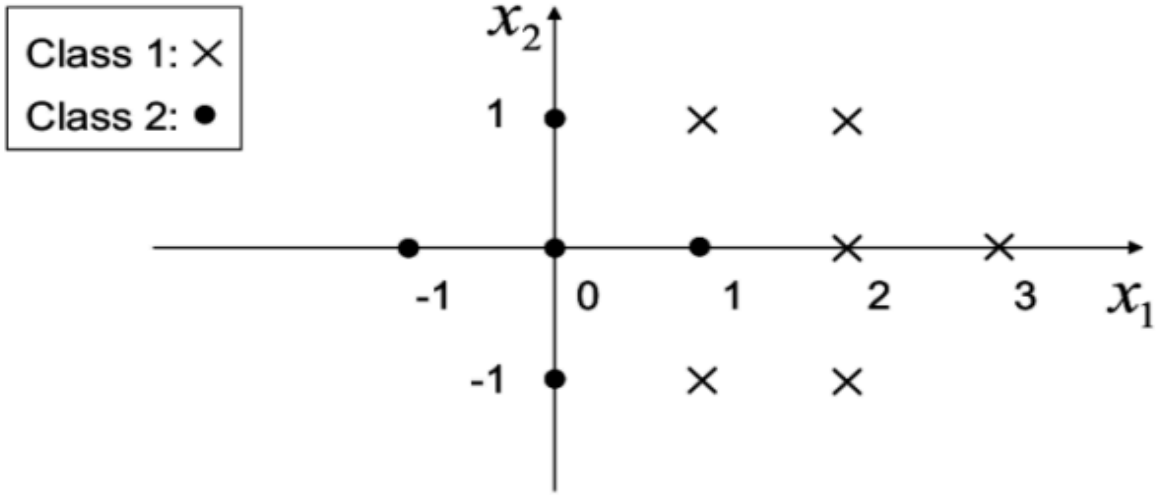
Accuracy = $\frac{TP+TN}{TP+FP+TN+FN} = \frac{300+200}{300+30+200+20} = 0.9091$

F1-score = $2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0.8531}{1.8466} = 0.9231$

KNN

سوال اول

در تصویر زیر تعدادي نمونه از دو کلاس مختلف مشخص شده اند. داده ي تست (0.5,0) را با روش KNN طبقه بندي کنید .



K نزدیک ترین همسایه با K=3 با دو فاصله ي زیر:

1. فاصله اقلیدسی
2. فاصله منهتن

Support Vector Machine

سوال اول

- به چه نقاطی support vector گفته می شود و آن را روی مثالی دلخواه نمایش دهید.
- به نظر شما طبقه بند SVM برای طبقه بندی چه نوع داده هایی مناسب نیستند؟
- درباره kernel ها و نقش آن ها در طبقه بندی توضیح دهید.(توضیح دهید وظیفه kernel ها چیست و چجوری به طبقه بندی کمک می کنند)
- تفاوت soft svm classifier با hard svm classifier بیان کنید.
- نحوه استفاده از SVM در مسائل رگرسیون رو را با کشیدن شکل توضیح دهید.