University of Tehran
دانشکدگان علوم و فناوری های میان رشته ای

Machine Learning

HW1

Fall 2024

توضیحات مهم

سلام بر دانشجویان عزیز، چند نکته مهم:

۱. حجم گزارش به هیچ عنوان معیار نمره‌دهی نیست، در حد نیاز توضیح دهید.

۲. نکته‌ی مهم در گزارش‌نویسی روشن بودن پاسخ‌ها می‌باشد، اگر فرضی برای حل سوال استفاده می‌کنید حتماً آن را ذکر کنید، اگر جواب نهایی عددی است به‌صورت واضح آن را بیان کنید.

۳. برای سوالات شبیه‌سازی، فقط از دیتاست داده‌شده استفاده کنید.

۴. فایل نهایی خود را در یک فایل زیپ شامل گزارش با فرمتPDF ، آپلود کنید. نام فایل زیپ ارسالی باید الگوی زیر را داشته باشد :

ML_HW1_StudentNumber

۵. کد سوالات شبیه سازی بصورت فایلی تحت عنوان سوال آن با پسوند .py یا .ipynb به همراه گزارش در فایل زیپ تحت الگوی بند ۴پیوست شود.

۶. هرگونه شباهت در گزارش و کد مربوط به شبیه‌سازی، به منزله‌ی تقلب می‌باشد و کل تمرین برای طرفین صرف‌نظر خواهد شد.

۷. تمامی سوالات خود را به دستیار آموزشی تمرین مربوطه به ایمیل زیر ارسال نمایید:

Mahsanadafi6@gmail.com

سوال اول:Maximum Likelihood

فرض کنید $\{x_k\}, k = 1,2, \dots, N$ نمونه‌های مستقل از یکی از توزیع‌های زیر هستند. حداکثر درست‌نمایی θ را برای هر کدام به دست آورید:

a) $f(x_k;\theta) = \theta \exp(-\theta x_k)\ x_k \geq 0, \theta > 0$ Exponential Density

c) $f(x_k;\theta) = \sqrt{\theta} x_k^{\sqrt{\theta}-1}\ 0 \leq x_k \leq 1, \theta > 0$ Beta Density

سوال دوم: Regression

الف. مرحله به مرحله رابطه رگرسیون خطی با استفاده از واریانس، میانگین، کواریانس و correlation طبق مراحل زیر اثبات کنید:

۱. مفاهیم اولیه واریانس، میانگین، کواریانس و correlation را تعریف کنید.

۲. در رگرسیون خطی ساده، هدف یافتن رابطه‌ای خطی بین متغیر وابسته Y و متغیر مستقل X است. مدل به صورت زیر تعریف می‌شود:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

که در آن:

- Yمتغیر وابسته (پاسخ) است.
- Xمتغیر مستقل (پیش‌بینی‌کننده) است.
- β_0عرض از مبداء و β_1 شیب است.
- ε جمله خطا است که تفاوت بین مقادیر مشاهده شده و پیش‌بینی شده Y را نشان می‌دهد.

با استفاده از رابطه بالا مجموع مربعات باقی‌مانده[1] را کمینه کنید.

۳. استخراج شیب (β_1) و عرض از مبدأ (β_0) بر حسب کوواریانس و واریانس

۴. چگونه می‌توانیم β_1 را بر حسب ضریب همبستگی (ρ) و انحراف معیارها بیان کنیم؟ این رابطه چه چیزی را درباره ارتباط بین شیب خط رگرسیون و قدرت همبستگی بین متغیرها نشان می‌دهد؟

۵. فرمول نهایی برای شیب (β_1) بر حسب کوواریانس و واریانس چیست؟ فرمول نهایی برای عرض از مبدأ (β_0) چگونه با میانگین‌های X و Y مرتبط است؟

[1] Sum of Squared Residuals(RSS)

۶. این روابط چه بینشی درباره ارتباط بین ضرایب رگرسیون و معیارهای آماری مانند واریانس، کوواریانس و همبستگی به ما می‌دهند؟ چگونه این روابط به ما در درک بهتر خط رگرسیون تخمین زده شده کمک می‌کنند؟

ب. جدول زیر مربوط به یک مسئله رگرسیون خطی ساده است یا استفاده از فرمول‌های بدست آمده در قسمت قبل واریانس، میانگین، کواریانس و correlation را گزارش کنید.(کد نویسی با پایتون آنجام شود.)

i	x_i	y_i
1	16	46
2	27	80
3	11	36
4	20	52
5	30	98
6	25	75
7	5	10
8	24	70
9	21	64
10	10	30

سوال سوم: طبقه بندی

در این سوال، یک طبقه‌بند طراحی کنید که بتواند دو کلاس متفاوت (دو تیم فوتبال منچستریونایتد و چلسی) را با استفاده از دیتاست داده‌شده تشخیص دهد. برای طبقه‌بندی، می‌توانید میانگین رنگ در هر عکس را محاسبه کنید و سپس مقدار به‌دست‌آمده را با رنگ‌های آبی و قرمز مقایسه نمایید. این طبقه‌بند را روی دیتاست داده‌شده تست کنید. ماتریس Confusion را گزارش دهید و مقادیر accuracy ، precision، و recall را محاسبه کرده و نتایج هر کدام را توضیح دهید.

نکته: بدین منظور میتوان از میانگین کانال های عکس RGB استفاده کرد.

سوال چهارم: بررسی برازش[1] توابع مختلف

ابتدا با توجه به کد زیر داده‌های مربوطه را تولید کنید:

```
x = np.arange(-10, 10, 0.2)
y = 2 * cos(x) / -pi + 2 * sin(2 * x) / (2 * pi) + 2 *
cos(3 * x) / (-3 * pi)
```

سپس داده‌ها را با استفاده از K-fold به k بخش تقسیم کنید.

الف. در ابتدا سعی کنید توابع زیر را برازش کنید و مقادیر میانگین MSE برای هر یک از موارد را گزارش دهید.

1. Linear Regression
2. Polynomial Regression
3. Ridge Regression (L2 Regularization)

ب. در این بخش ضریب regularization رو در ridge Regression را بهینه و گزارش کنید. (در این بخش رسم نمودار MSE در فرایند بهینه سازی نمره اضافه به همراه دارد.)

[1] Fitting