شیوه تحویل تمرینات

تمرینات نوشتاری: تحویل در کلاس درس (زمان تحویل: ۲۸ مهر تا انتهای زمان کلاس)
تمرینات کامپیوتری: آپلود در سایت quera (زمان تحویل: ۲۷ مهر تا ساعت ۲۳:۵۹ دقیقه)

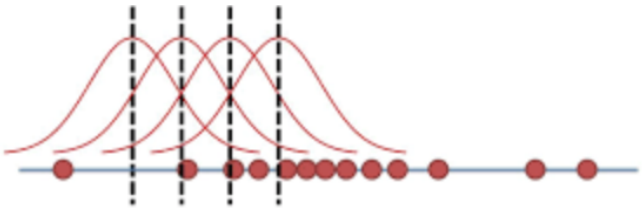
فایل های فرستاده شده باید شامل گزارش در قالب فایل pdf و کدها با پسوند py. باشد.

تذکر: در تمرینات کامپیوتری سهم عمده نمره را **تحلیل** و **دریافت** شما از نتایج کدها دارد. سعی کنید گزارش خلاصه و شامل نکات مهم بوده و خروجی کدها در آن آورده شود.

تمرین‌های نوشتاری

سوال ۱ : مقدمات احتمال

الف) بیشینه امکان یا Maximum Likelihood، مفهومی برای محاسبه‌ی پارامترهای یک مدل احتمالاتی است. برای شفاف تر شدن مفهوم آن، شکل زیر را در نظر بگیرید:



به عنوان مثال اگر بدانیم داده‌ها از یک توزیع گاوسی با یک واریانس معلوم آمده‌اند و به دنبال مرکز این توزیع هستیم، در این روش، پارامتر توزیع (مرکز آن) را عددی می گیریم که احتمال تولید شدن این داده‌ها از آن، بیش از سایر توزیع ها باشد. فرمول بندی ریاضی آن بصورت زیر است:

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}}[\mathbb{P}(X|\theta)]$$

که در اینجا، اگر داده‌ها بصورت مستقل و هم توزیع (i.i.d) انتخاب شده باشند، داریم:

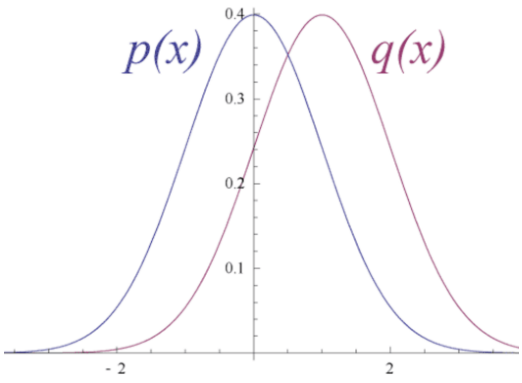
$$\mathbb{P}(X|\theta) = \prod_i \mathbb{P}(x_i|\theta)$$

به دلیل سادگی، در بسیاری از اوقات تابع Log-Likelihood را بیشینه می کنیم که معادل با بیشینه کردن خود Likelihood است:

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}}[\log \mathbb{P}(X|\theta)] = \underset{\theta}{\operatorname{argmax}}[\sum_i \log \mathbb{P}(x_i|\theta)]$$

- فرض کنید n داده i.i.d به صورت $x_1, x_2, \ldots, x_n$ داریم که می دانیم از یک توزیع گاوسی با واریانس مشخص σ و میانگین مجهول (μ_i) داده شده اند (داده‌ها یک بعدی اند). تخمین Maximum Likelihood از میانگین، μ_{ML} را بدست آورید.

برای تعیین شباهت یا تفاوت دو توزیع، معیارهای مختلفی تعریف می شود که به فواصل بین دو توزیع معروف اند. یکی از این فواصل، فاصله‌ی Kullback-Leibler یا به طور خلاصه فاصله‌ی KL است که به صورت زیر تعریف می‌شود:



$$D_{KL}[p(x) \,||\, q(x)] = \mathbb{E}_{x\sim p(x)}[log\frac{p(x)}{q(x)}] = \int_x p(x)\, log\frac{p(x)}{q(x)}\, dx$$

که در حالت گسسته، انتگرال بالا به جمع تبدیل می شود. همانطور که مشخص است فاصله‌ی دو توزیع یکسان، صفر است. این فاصله مثبت است و کران بالا ندارد.

ب) میتوان انتگرال بالا را به صورت زیر باز کرد:

$$\int_x p(x)\, log\frac{p(x)}{q(x)}\, dx = \int_x p(x)\, log[p(x)]\, dx - \int_x p(x)\, log[q(x)]\, dx$$

توضیح دهید اگر $p(x)$ توزیع واقعی داده‌ها و $q(x)$ توزیعی (مثلا با پارامتر θ)باشد که میخواهیم آن را پیدا کنیم، چرا کمینه کردن فاصله‌ی KL میان دو توزیع، معادل با بیشینه کردن Maximum Likelihood است.

پ) در حالت یک بعدی (x در فضای یک بعدی است)، فاصله KL بین دو توزیع گاوسی $\mathcal{N}(\mu_1, \sigma_1)$و $\mathcal{N}(\mu_2, \sigma_2)$ را برحسب این پارامترها بدست آورید.

راهنمایی:

$$\begin{aligned} \mathrm{E}_X[(X-\mu)^2] &= \int_{-\infty}^{\infty} (x-\mu)^2 p(x;\mu,\sigma^2)dx \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})dx \\ &= \sigma^2 \end{aligned}$$

تمرین‌های کامپیوتری

سوال ۱: آشنایی با پایتون

الف) با استفاده از بخش random کتابخانه numpy مجموعه‌ای ۱۰۰۰۰ نمونه‌ای از اعداد تصادفی با توزیع گاوسی با میانگین ۱۰ و واریانس ۱ تولید نمایید.

ب) هیستوگرام اعداد تولید شده در بخش پیشین را توسط کتابخانه matplotlib رسم نمایید.

پ) با استفاده از دستور sample دویست نمونه تصادفی از بردار اول را انتخاب نمایید.

ت) هیستوگرام اعداد نمونه‌برداری شده را رسم نموده و در مورد تفاوت آن با هیستوگرام بخش ب بحث کنید.

ث) یک مجموعه ۱۰۰۰۰ نمونه‌ای از اعداد تصادفی با توزیع یکنواخت بین ۱- و ۱ تولید کرده و هیستوگرام آن را رسم کنید.

ج) مجموعه اعداد بخش الف و ث را در هم ضرب کرده و هیستوگرام حاصل را رسم کنید. در مورد توزیع این دادگان بحث کنید.

سوال ۲: آشنایی با پردازش تصاویر

الف) با استفاده از کتابخانه pandas مجموعه اعداد فایل img.csv را خوانده و در یک ماتریس به فرمت numpy بریزید

ب) با استفاده از کتابخانه matplotlib ماتریس مورد نظر را به صورت تصویر نمایش داده و در گزارش خود بیاورید.

پ) با استفاده از کتابخانه opencv-cv2 تصویر img.jpg را لود کنید.

ت) با استفاده از کتابخانه cv2 تصویر مورد نظر را نمایش دهید

ث) تصویر مورد نظر را با استفاده از فیلتر میانگین به ابعاد ۱۰ فیلتر کنید.

ج) خروجی مورد نظر را در فایل output.jpg ذخیره نموده و در گزارش بیاورید.

سوال ۳: آشنایی با کتابخانه یادگیری ماشین

الف) در این بخش می خواهیم اثر underfit و overfit را بسته به پیچیدگی مدل مشاهده کنیم و با اندازه گیری دقت مدل بیشتر آشنا شویم. دیتاست را اینگونه لود کنید:

```
from sklearn.datasets import load_digits
digits = load_digits()
```

که ۱۷۹۷ تصویر ۸ در ۸ از اعداد دست نویس را در اختیار شما قرار می دهد. digits.data، بردار پیکسل های تصاویر و digits.target بردار عدد آنها ((Labelمی باشد.

ب) به طور نمونه، تصویر اعداد ۰ تا ۹ از این مجموعه را با imshow در matplotlib.pyplot رسم کنید.

به صورت زیر می توانید درصدی از داده‌ها را برای تست جدا کنید:

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(digits.data, digits.target, test_size=0.25, random_state=0)
```

با استفاده از sklearn.svm، می توانید به صورت زیر با polynomial svm دسته بندی کنید:

```
svmClassifier = svm.SVC(gamma=0.01, kernel='poly', degree=degree)
```

که در این شی، .fit، مدل را آموزش می دهد، .predict، خروجی مدل آموزش داده به ازای ورودی های تست (یا ولیدیشن) را می دهد و .score، روی گروهی از ورودی ها اعمال شده و دقت مدل را خروجی می دهد.

پ) درجه‌ی چند جمله‌ای در polynomial svm را از ۱ تا ۱۰ تغییر دهید و نمودار دقت مدل برحسب درجه را روی داده‌های تست رسم کنید. چه مشاهده می کنید؟

ت) فرض کنید میخواهیم عدد ۱را شناسایی کنیم؛ یعنی داده‌های نشانگر عدد ۱، Positive و بقیه Negative باشند. آنگاه مدل ما اگر ۱ را به درستی ۱ تشخیص دهد True Positive است و برای سه حالت دیگر نیز مشابها بدست می آید. برای مدلی با بهترین درجه (درجه‌ای که دقت مدل با توجه به نتیجه‌ی قسمت قبل بیشینه بوده است) معیارهای Sensitivity, Specificity, Precision, Accuracy و Dice را محاسبه کنید.