به نام خدا

تمرین سری دوم **یادگیری آماری** نیم‌سال اول ۰۳-۰۲

---

۱. (۲۰ نمره) پاسخ کوتاه

به سوالات زیر به صورت کوتاه پاسخ دهید:

- تفاوت بین forward selection و backward selection را برای انتخاب متغیرها توضیح دهید. آیا به ازای تعداد مشخصی متغیر مورد نظر، مجموعه متغیرهای این دو روش یکسان خواهد شد؟
- شما در حال طراحی مدلی بر روی یک دیتاست با تعداد ۱۰۰۰ ویژگی برای یک تسک رگرسیون (regression) هستید. در ابتدا مدل خود را بر روی ۱۰۰ نمونه آموزش می‌دهید و مشاهده می‌کنید که با وجود همگرا شدن آموزش، خطای آموزش بر روی این نمونه‌ها زیاد است. پس در ادامه تصمیم می‌گیرید که شبکه خود را این‌بار روی ۱۰۰۰۰ نمونه آموزش دهید. آیا روش شما برای حل این مشکل صحیح است؟ اگر بلی، محتمل‌ترین نتایج مدل خود را در این حالت توضیح دهید. اگر خیر، راه‌حلی برای رفع این مشکل بیان کنید.
- هر چه بردارهای ویژه‌ای از ماتریس کواریانس که برای کاهش ابعاد از طریق PCA استفاده می‌کنیم دارای مقدار ویژه‌ی بزرگتری باشند، خطای بازسازی کمتر می‌شود. دلیل این موضوع را به صورت خلاصه توضیح دهید.
- خطای روی داده‌های آموزش و تست را در دو حالت overfitting و underfitting مقایسه کنید.

۲. (۳۵ نمره) رگرسیون خطی، تخمین ML و تخمین MAP

همانطور که از درس می‌دانید، در یک مدل رگرسیون خطی با ویژگی‌های x_i داریم:

$$y = \sum_{i=1}^{p} w_i x_i + \epsilon = w^T x + \epsilon$$

در صورتی که نویز موجود دارای توزیع $\epsilon \sim \mathcal{N}(0, \sigma^2)$ باشد، مشخصا خواهیم داشت:

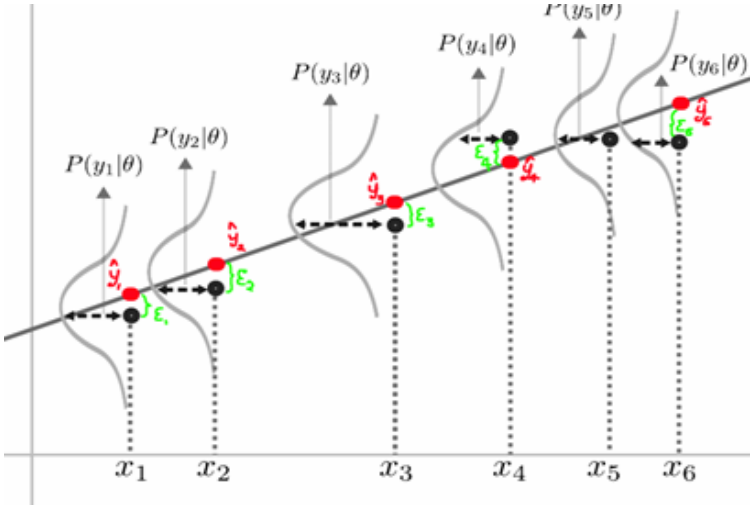
$$y|x, w \sim \mathcal{N}(w^T x, \sigma^2)$$

با در نظر گرفتن تمام نمونه‌های آموزشی می‌توان این عبارت را برای همه آن‌ها بنویسیم و در نتیجه به صورت برداری خواهیم داشت:

$$Y|X, w \sim \mathcal{N}(Xw, \sigma^2 I_n)$$

که در عبارت بالا $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ و $w \in \mathbb{R}^p$ می‌باشد.

الف) توزیع بالا به چه معناست؟ برای راهنمایی می‌توانید از شکل زیر کمک بگیرید.



نکته: در ۳ بخش بعدی جواب خود را به صورت یک مسئله بهینه‌سازی کمترین مربعات (که می‌تواند همراه با یک جمله regularizer باشد) بنویسید و نیازی به محاسبه $\hat{w}_{MAP}$ و $\hat{w}_{ML}$ نیست.

ب) تخمین ML را برای w بدست بیاورید.
این مسئله معادل با کدام حالت روش رگرسیون است؟

پ) فرض کنید برای پارامترهای w یک توزیع اولیه (Prior) در نظر می‌گیریم؛ به طوریکه $w \sim \mathcal{N}(۰, \lambda^۲ I_p)$.
تخمین MAP را برای w بدست آورید.
این مسئله معادل با کدام حالت روش رگرسیون است؟

ج) حال توزیع اولیه را تغییر می‌دهیم. فرض کنید که هر یک از وزن‌ها دارای توزیع $w_i \sim Laplace(۰, \lambda)$ باشند. تخمین MAP را برای w بدست آورید.
این مسئله معادل با کدام حالت روش رگرسیون است؟

چ) تفاوت بین استفاده از این دو توزیع را از دیدگاه اثر آن‌ها بر روی اندازه w_i ها به صورت خلاصه توضیح دهید.

راهنمایی:

$$Z \sim Laplace(۰, \lambda) \rightarrow f_Z(z) = \frac{۱}{۲\lambda} exp(-\frac{|z|}{\lambda})$$

$$Z \sim \mathcal{N}(\mu, \Sigma) \rightarrow f_Z(\mathbf{z}) = \frac{۱}{(۲\pi)^{n/۲}|\Sigma|^{۱/۲}} exp(-\frac{۱}{۲}(\mathbf{z}-\mu)^T \Sigma^{-۱}(\mathbf{z}-\mu))$$

۳. (۲۵ نمره) رگرسیون خطی Ridge

مدل رگرسیون خطی $y = X\beta + \epsilon$ که $X \in \mathbb{R}^{n \times p}$ با کمترین مربعات رگولایز شده $L_۲$ را در نظر بگیرید:

$$\min_{\beta} \|y - X\beta\|_۲^۲ + \lambda \|\beta\|_۲^۲$$

که λ پارامتر رگولاریزاسیون است.

الف) فرم بسته $\hat{\beta}^{ridge}(\lambda)$ را بدست آورید.

ب) اگر $\epsilon \in \mathcal{N}(۰, \sigma^۲ I_n)$ باشد، ثابت کنید که کوواریانس $\hat{\beta}^{ridge}(\lambda)$ به شکل زیر است:

$$Cov(\hat{\beta}^{ridge}(\lambda)) = \sigma^۲(X^T X + \lambda I)^{-۱} X^T X (X^T X + \lambda I)^{-۱}$$

پ) اگر ماتریس X متعامد یکه (Orthonormal) باشد، رابطه بین $\hat{\beta}_j^{ridge}(\lambda)$ و $\hat{\beta}_j^{LS}(\lambda)$ (تخمین‌گر کمترین مربعات معمولی) را برای $j = ۱, ۲, ..., p$ بدست بیاورید. چه انتخابی برای λ، مقدار $\|\hat{\beta}^{ridge}(\lambda)\|_۲^۲$ را نصف مقدار $\|\hat{\beta}^{LS}(\lambda)\|_۲^۲$ می‌کند؟

۴. (۲۰ نمره) خطای بازسازی PCA

می‌خواهیم عمل PCA را انجام دهیم. هر نمونه $x_i \in \mathbb{R}^p$ به $z_i = V_{۱:k}^T x_i$ تصویر می‌شود. در اینجا $V_{۱:k}$ در واقع $[v_۱|v_۲|...|v_k]$ یا به عبارتی دیگر همان k مولفه اساسی اول است. می‌توانیم x_i را از روی z_i با استفاده از رابطه $\hat{x}_i = V_{۱:k} z_i$ بازسازی نماییم.

الف) نشان دهید

$$\|\hat{x}_i - \hat{x}_j\|_۲ = \|z_i - z_j\|_۲$$

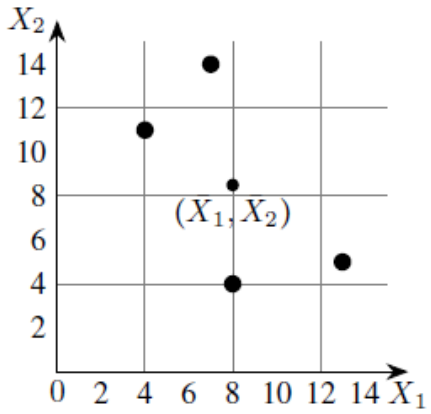
ب) نشان دهید خطای بازسازی برابر است با:

$$\sum_{i=۱}^{n} \|x_i - \hat{x}_i\|_۲^۲ = (n-۱) \sum_{i=k+۱}^{p} \lambda_i$$

چه برداشتی از این معادله راجع به خطای بازسازی می‌توان انجام داد؟

۵. (۱۰ نمره) انجام PCA به صورت دستی!

مجموعه دادگان زیر را در نظر بگیرید:



مولفه اساسی اول را پیدا و داده‌ها در راستای آن تصویر کنید.

۶. (۵۰ نمره) بخش عملی

ابتدا دیتاست بوستون را با استفاده از دستور زیر لود کنید:

```
from sklearn.datasets import load_boston
Boston = load_boston()
```

سپس با استفاده از دستور توضیحات(print(Boston.DESCR لازم درباره دیتاست را مطالعه نمایید. در این مسئله می‌خواهیم با استفاده از ویژگی‌هایی که در دیتاست آمده است، میانگین قیمت خانه را تخمین بزنیم.

الف) ابتدا داده‌ها را به نسبت 0.7 به 0.3 به داده‌های آموزش و تست تقسیم نمایید و سپس با استفاده از تمام ویژگی‌ها و اعمال رگرسیون خطی مقدار ضرائب را گزارش کنید و همچنین مقدار MSE و R^2 هم برای داده‌های آموزش و هم داده‌های تست گزارش نمایید.

ب) با استفاده از روش Forward Slection، سه ویژگی برتر را استخراج نمائید و سپس تنها با این سه ویژگی رگرسیون خطی را اعمال نمائید و مقدار MSE ، R^2 و همچنین ضرائب را گزارش کنید و با خروجی‌های بالا مقایسه نمائید. پ) حال مشخص نمایید که هر کدام از مولفه‌های اساسی داده‌های آموزش، چه بخشی از واریانس

را شامل می‌شود (می‌توانید explained variance ratio را بررسی نمایید) و آن را در یک نمودار ترسیم کنید (مولفه اول بیشترین واریانس و مولفه آخر کمترین واریانس). سپس سه مولفه اول را انتخاب نمایید و داده‌ها را در این سه راستا تصویر کنید و نهایتا رگرسیون خطی را به این سه ویژگی جدیدی که استخراج نموده‌اید اعمال کنید و مقدار MSE ، R^2 و همچنین ضرائب را گزارش کنید و با خروجی‌های بالا مقایسه نمائید.

ت) در این قسمت می‌خواهیم از روش Ridge Regression استفاده نماییم. برای یکسان سازی نتایج، پارامتر آلفا را به صورت زیر قرار دهید:

```
alphas = 10**np.linspace(3,-3,100)*0.5
```

سپس نمودارهای زیر را رسم نمایید:

- مقدار MSE برای داده‌های آموزش و تست بر حسب آلفا
- ضرائب بر حسب آلفا
- نمودار تعداد ویژگی‌های حذف شده بر حسب آلفا

نهایتا مقدار بهینه آلفا را انتخاب نمایید و مقدار MSE ، R^2 و همچنین ضرائب را گزارش کنید و با خروجی‌های روش رگرسیون معمولی مقایسه نمایید.

ث) قسمت د را برای روش Lasso Regression تکرار نمایید. به نظر شما در چه جاهایی نیاز است تا از روش Lasso استفاده نماییم؟

ج) در این قسمت می‌خواهیم بررسی کنیم که چگونه روش‌های Ridge و Lasso با بیش‌برازشی مقابله می‌کنند. برای اینکار نسبت داده‌های آموزشی به کل داده‌ها را از کم به زیاد تغییر دهید و هر چقدر این نسبت کمتر باشد، خطر بیش‌برازشی بیشتر خواهد شد. حال به ازای هر نسبت، هر سه مدل رگرسیون خطی، رگرسیون Ridge و رگرسیون Lasso را به داده‌ها فیت نمایید و سپس نمودارهای زیر را رسم کنید:

- مقدار R^2 برای داده‌های تست بر حسب نسبت داده‌های آموزشی به کل داده‌ها (هر سه روش در یک نمودار)
- مقدار آلفا انتخاب‌شده بر حسب نسبت داده‌های آموزشی به کل داده‌ها برای روش‌های Ridge و Lasso (در یک نمودار)

روند کلی این نمودارها به چه صورت است و این روند را به چه صورت تحلیل می‌کنید؟