Homework 2

❗ For each problem, please clearly show your reasoning and write all the steps.
G As data scientists, you should feel free to google it whenever you see something unfamiliar.
☺ Group discussion for the homework is encouraged, but you have to write your answer by yourself. Also, you are always welcome to discuss the problems with me.

Task 0.

Read the relevant chapters in the textbooks listed on Sakai.

Problem 1. Polynomial regression (5pt)

Consider the polynomial regression problem with

$$g(x_n|w_0, w_1, \cdots, w_k) = \sum_{l=0}^{k} w_l (x_n)^l = w_k (x_n)^k + \cdots + w_2 (x_n)^2 + w_1 x_n + w_0$$

and recall that the error function is given by

$$E(w_0, w_1, \cdots, w_k|\mathcal{X}) = \frac{1}{2} \sum_{n=1}^{N} \left(y_n - g(x_n|w_0, w_1, \cdots, w_k)\right)^2 \ .$$

1. By minimizing the above error function, show that the maximum likelihood solution for $w_0, w_1, \cdots, w_k$ is given by the solution of $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{y}$, where $\boldsymbol{w} = [w_0, \cdots, w_k]^{\mathrm{T}} \in \mathbb{R}^{k+1}$ is the vector of paramters,

$$\boldsymbol{A} = \begin{bmatrix} N & \sum_{n=1}^N x_n & \sum_{n=1}^N (x_n)^2 & \cdots & \sum_{n=1}^N (x_n)^k \\ \sum_{n=1}^N x_n & \sum_{n=1}^N (x_n)^2 & \sum_{n=1}^N (x_n)^3 & \cdots & \sum_{n=1}^N (x_n)^{k+1} \\ \vdots & \vdots & \ddots & & \\ \sum_{n=1}^N (x_n)^k & \sum_{n=1}^N (x_n)^{k+1} & \sum_{n=1}^N (x_n)^{k+2} & \cdots & \sum_{n=1}^N (x_n)^{2k} \end{bmatrix} \in \mathbb{R}^{(k+1)\times(k+1)} \ ,$$

and

$$\boldsymbol{r} = \begin{bmatrix} \sum_{n=1}^N y_n \\ \sum_{n=1}^N y_n x_n \\ \sum_{n=1}^N y_n (x_n)^2 \\ \vdots \\ \sum_{n=1}^N y_n (x_n)^k \end{bmatrix} \in \mathbb{R}^{k+1} \ .$$

2. Notice that $\boldsymbol{A} = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}$ and $\boldsymbol{r} = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{y}$ where

$$\boldsymbol{D} = \begin{bmatrix} 1 & x_1 & (x_1)^2 & \cdots & (x_1)^k \\ 1 & x_2 & (x_2)^2 & \cdots & (x_2)^k \\ \vdots & \vdots & \ddots & & \\ 1 & x_N & (x_N)^2 & \cdots & (x_N)^k \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \ .$$

Under what condition can we say

$$\boldsymbol{w} = (\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})^{-1}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{r} \ ?$$

That is, when is $\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}$ invertible? What does it mean in terms of the sample?

Problem 2. LASSO (10pt)

We mentioned that LASSO stands for "**l**east **a**bsolute **s**hrinkage and **s**election **o**perator" in class. Now you kind of understand it shrinks the parameters by selecting important features and that is how it got its name. (On the other hand, "ridge" describes the fact that you add a diagonal $\lambda\boldsymbol{I}$ to the original matrix which looks like a "ridge" in the matrix.)

This problem helps you understand the regime of LASSO. It is adapted from Ex. 3.27 in [HaTF]. It is highly recommended that you work your own way through the problem instead of copying from the solution manual which can be easily found at this link. Even if you copy the solution from the solution manual, you will not get full credit because their solution is wrong.

Consider the LASSO problem in Lagrange multiplier form with $L(\boldsymbol{w}) = \sum_n (r_n - \sum_j x_{nj} w_j)^2$ (where $x_{n0} = 1$ for all t) and we want to solve

$$\min_{\boldsymbol{w}} \ L(\boldsymbol{w}) + \lambda \sum_j |w_j| \qquad (\clubsuit)$$

for a fixed hyperparamater $\lambda > 0$.

1. Setting $w_j = w_j^+ - w_j^-$ with $w_j^+, w_j^- \geq 0$, ($\clubsuit$) becomes $L(\boldsymbol{w}) + \lambda \sum_j (w_j^+ + w_j^-)$. Show that the Lagrange dual function is

$$L(\boldsymbol{w}) + \lambda \sum_j (w_j^+ + w_j^-) - \sum_j \lambda_j^+ w_j^+ - \sum_j \lambda_j^- w_j^-$$

and the KKT optimality conditions are

$$\begin{aligned} \nabla L(\boldsymbol{w})_j + \lambda - \lambda_j^+ &= 0 \\ -\nabla L(\boldsymbol{w})_j + \lambda - \lambda_j^- &= 0 \\ \lambda_j^+ w_j^+ &= 0 \\ \lambda_j^- w_j^- &= 0 \end{aligned}$$

along with the non-negativity constraints on the parameters and all the Lagrange multipliers.

2. Show that $|\nabla L(\boldsymbol{w})_j| \leq \lambda$ for all j and that the KKT conditions imply the following three scenarios:

$$\begin{aligned} \lambda = 0 \ &\Rightarrow \ \nabla L(\boldsymbol{w})_j = 0 \ \text{ for all } j \\ w_j^+ > 0, \lambda > 0 \ &\Rightarrow \ \lambda_j^+ = 0, \nabla L(\boldsymbol{w})_j = -\lambda < 0, w_j^- = 0 \\ w_j^- > 0, \lambda > 0 \ &\Rightarrow \ \lambda_j^- = 0, \nabla L(\boldsymbol{w})_j = \lambda > 0, w_j^+ = 0 \end{aligned}$$

and hence show that for any "active"predictor having $w_j \neq 0$, we must have $\nabla L(\boldsymbol{w})_j = -\lambda$ if $w_j > 0$, and $\nabla L(\boldsymbol{w})_j = \lambda$ if $w_j < 0$. Assuming the predictors (the independent variables) are standardized, relate λ to the correlation between the j-th predictor $\boldsymbol{x}_j$ and the current residuals $(\boldsymbol{r} - \boldsymbol{X}\boldsymbol{w})$).

3. Now let the hyperparameter λ vary. The solution $\hat{\boldsymbol{w}}_{\text{LASSO}}$ can be treated as a function of λ. Suppose the set of active predictors is unchanged for $\lambda_0 \geq \lambda \geq \lambda_1$. Show that there is a vector $\boldsymbol{\gamma}_0$ such that

$$\hat{\boldsymbol{w}}_{\text{LASSO}}(\lambda) = \hat{\boldsymbol{w}}_{\text{LASSO}}(\lambda_0) - (\lambda - \lambda_0)\boldsymbol{\gamma}_0$$

and thus the LASSO solution path in linear as λ ranges from λ_0 to λ_1.

Problem 3. Bias-(co)variance calculation for ridge regression (10pt)

Consider the model $\boldsymbol{r} = \boldsymbol{X}\boldsymbol{w}_* + \epsilon$ where $\boldsymbol{r} \in \mathbb{R}^N$. Suppose $\boldsymbol{X}$ is known and fixed (so that the only randomness in your sample $\{(\boldsymbol{x}_n, r_n)\}_{n=1}^N$ comes from ϵ. Suppose $\epsilon \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

1. Suppose we want to estimate $\boldsymbol{w}_*$ using the ridge regression with a hyperparameter λ. Write $\hat{\boldsymbol{w}}_{\text{ridge}}$ in terms of $\boldsymbol{X}$, λ, $\boldsymbol{w}_*$ and ϵ.
2. Find $\mathbb{E}_{\epsilon\sim\mathcal{N}(\boldsymbol{0},\sigma^2\boldsymbol{I})}[\hat{\boldsymbol{w}}_{\text{ridge}}]$. Is the estimator biased for $\boldsymbol{w}_*$?
3. Find $\text{Cov}(\hat{\boldsymbol{w}}_{\text{ridge}})$, the covariance matrix for the random vector $\hat{\boldsymbol{w}}_{\text{ridge}}$.
4. The nuclear norm $\|\boldsymbol{A}\|_{(1)}$ of a covariance matrix $\boldsymbol{A}$ is equal to its trace $\text{tr}\boldsymbol{A}$ (not true for a general matrix). Calculate $\|\text{Cov}(\hat{\boldsymbol{w}}_{\text{ridge}})\|_{(1)}$.
5. Discuss, in plain language, how λ affects the bias-covariance decomposition. You can use the nuclear norm as a measure of significance for the covariance matrix.

Problem 4. Nonparametric density estimation (5pt)

This problem is [Bi] Ex.2.60. Please think about the problem carefully before you (if you really wish to) consult with the solution manual of [Bi] at this link.

Consider a histogram-like density model in which the space is divided into fixed regions for which the density $p(\boldsymbol{x})$ takes the constant value h_i over the i-th region, and that the volume of region i is denoted Δ_i. Suppose we have a set of N observations of $\boldsymbol{x}$ such that n_i of these observations fall in region i. Using a Lagrange multiplier to enforce the normalization constraint on the density, derive an expression for the maximum likelihood estimator for the $\{h_i\}$. Here $\{h_i\}$ are considered as the parameters of the density estimator.

Problem 5. EM for Bernoulli (5pt)

Consider a random variable $\boldsymbol{x} \in \mathbb{R}^D$ with binary elements. A multivariate Bernoulli mixture model is given by

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \pi) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k) \ ,$$

where

$$p(\boldsymbol{x}|\boldsymbol{\mu}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{(1-x_i)} \ .$$

Derive the EM algorithm for the multivariate Bernoulli mixture model where the data points are in $\mathbb{R}^D$. Clearly show all your steps.

Problem 6. Programming: EM-GMM vs K-means (5pt + 2 bonus pts for nice pictures)

Do Python for the following tasks. (If you have a strong preference, you can also use other coding languages.) Your codes have to execute the algorithm we discussed in class and you are NOT allowed to call a K-means function/class or an EM function/class from a library.

1. In $\mathbb{R}^2$, from a Gaussian mixture model (GMM) with $K = 2$ components, generate 100 points. The GMM $p(\boldsymbol{x}) = \sum_{k=1}^{2} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ satisfies $\pi_1 = \pi_2 = 0.5$, and $\boldsymbol{\mu}_1 = (0, 8)^{\mathrm{T}}, \boldsymbol{\mu}_2 = (2, 4)^{\mathrm{T}}, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$. Add a noise $\sim \mathcal{N}(0, 0.25^2 \boldsymbol{I})$ to each point. Plot the data points and use different colors to imply their clusters.
2. Perform K-means with $K = 2$ to the data. Show your results in color.
3. Perform EM with $K = 2$ to the data. Show your results in color.

Hints for Problem 2, Part 3

Note that from Part 2, we have

$$\lambda \text{sign}(w_j) = 2\boldsymbol{x}_j^{\mathrm{T}}(\boldsymbol{r} - \boldsymbol{X}\boldsymbol{w})$$

where $\boldsymbol{x}_j$ denotes the j-th column of $\boldsymbol{x}$.
Let $\tilde{\boldsymbol{w}}$ denote the vector of active w_j's. We can definitely create an matrix $\tilde{\boldsymbol{X}}$ which only consists of columns of $\boldsymbol{X}$ that are "active". Specifically, let $\mathcal{J}$ denotes the set of active indices. $\tilde{\boldsymbol{w}} = (w_j)_{j\in\mathcal{J}}$ and $\tilde{\boldsymbol{X}} = [\boldsymbol{x}_j]_{j\in\mathcal{J}}$.
Then since only w_j's for $j \in \mathcal{J}$ counts, we have

$$\lambda \text{sign}(w_j) = 2\boldsymbol{x}_j^{\mathrm{T}}(\boldsymbol{r} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{w}}) \ .$$

Now define a vector $\boldsymbol{s}$ whose length is $|\mathcal{J}|$ and j-th entry (corresponding to the index of $\boldsymbol{w}$) is

$$s_j = \text{sign}(w_j) \ .$$

It is clear that $\boldsymbol{s}$ does not change when λ moves from λ_0 to λ_1 (the sign cannot jump from + to − without going through 0 by continuity). So we can treat $\boldsymbol{s}$ as a constant vector when we change λ.
Now we have, to emphasize the dependence on λ,

$$\lambda\boldsymbol{s} = 2\tilde{\boldsymbol{X}}^{\mathrm{T}}(\boldsymbol{r} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{w}}(\lambda)) \ .$$

Also,

$$\lambda_0\boldsymbol{s} = 2\tilde{\boldsymbol{X}}^{\mathrm{T}}(\boldsymbol{r} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{w}}(\lambda_0)) \ .$$

Taking the difference,

$$(\lambda - \lambda_0)\boldsymbol{s} = -2\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}}\left[\tilde{\boldsymbol{w}}(\lambda) - \tilde{\boldsymbol{w}}(\lambda_0)\right] \ .$$

That is,

$$\tilde{\boldsymbol{w}}(\lambda) - \tilde{\boldsymbol{w}}(\lambda_0) = (\lambda - \lambda_0)\left[-\frac{1}{2}(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}})^{-1}\boldsymbol{s}\right] = (\lambda - \lambda_0)\tilde{\gamma}_0 \ ,$$

where $\tilde{\gamma}_0 = -\frac{1}{2}(\tilde{\boldsymbol{X}}^{\mathrm{T}}\tilde{\boldsymbol{X}})^{-1}\boldsymbol{s}$. Now plug in zero entries to make a vector γ_0 where the j-th entry of γ_0 is equal to the corresponding entry of γ if $j \in \mathcal{J}$, but equal to 0 if $j \notin \mathcal{J}$. Note that for $j \notin \mathcal{J}$, $w_j(\lambda) - w_j(\lambda_0) = 0 - 0 = 0 = (\lambda - \lambda_0) \cdot 0$. What can we conclude?