Midterm Exam #1

COEN 240 Winter 2021

Q1. (25 pts) Multi-select questions (5×3 pts) and short answer questions (2×5 pts). For each multi-select question, we will dock 2 pts for incomplete selections and 3 pts if there are any incorrect choices.

(1) [AD] Please select all the supervised learning tasks from the following.

(A) Classification
(B) Density estimation
(C) Stochastic Gradient Descent
(D) Regression

(2) [BD] Please select all the possible reasons leading to the model underfitting issue.

(A) The model is optimized by the Gradient Descent algorithm with a given budget (*i.e.*, max iteration steps) and the learning rate has been set too small.
(B) The hyperparameter controlling the regularization term has been set too small.
(C) The hyperparameter controlling the regularization term has been set too large.
(D) The model capacity is insufficient, *e.g.,* the order M of a polynomial regression function is too small (Polynomial regression: $y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j$).

(3) [AC] Please select all the possible solutions to alleviate the model overfitting issue.

(A) Add a regularization term into the loss function.
(B) Increase the training data size.
(C) Tune the hyperparameters controlling the regularization term on the testing dataset.
(D) Increase the model complexity, *i.e.*, by adding more model parameters (*e.g.,* Increase M in our polynomial curve fitting example).

(4) [AC] Please select all the linear regression models from the following.

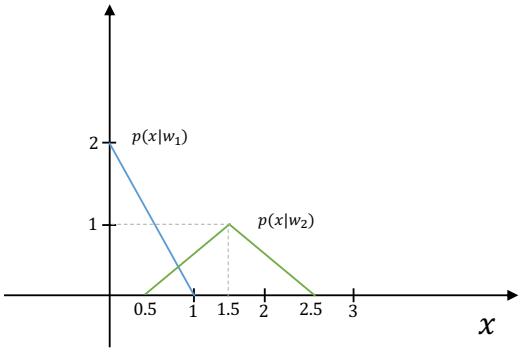
(A) Ridge regression.
(B) Softmax regression.
(C) Least Absolute Shrinkage and Selection Operator (LASSO)
(D) Logistic regression.

(5) [BC] Please select all the limitations for a Naive Bayesian (NB) classifier.

(A) The NB method can only be used with discrete features.
(B) The NB method may suffer from a zero-frequency issue.
(C) The NB method has to assume that all the features are statistically independent.
(D) The NB method has to model the density function for the features of data samples.

(6) Please give the Gradient Descent (GD) solution for a Ridge regression model defined with the loss by $\mathcal{L}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y_n - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$, where $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{x}_n \in \mathbb{R}^d$ (Hint: first compute the gradient of $\mathcal{L}(\mathbf{w})$ w.r.t $\mathbf{w}$, then show the iterative update as a pseudo algorithm). Can GD lead to a global optima for Ridge regression? Why?

(7) Please point out the differences between MLE, MAP, and full Bayesian parameter estimation.



$$p(x|w_1) = \begin{cases} -2x + 2, & 0 \ll x \ll 1 \\ 0\ , & else \end{cases}$$

$$p(x|w_2) = \begin{cases} x - \frac{1}{2}, & \frac{1}{2} \ll x \ll \frac{3}{2} \\ -x + \frac{5}{2}, & \frac{3}{2} \ll x \ll \frac{5}{2} \\ 0, & else \end{cases}$$

Figure 1: Problem setting of Q2.

Q2. (30 pts) Bayes Decision Theory: a) 5 pts, b) 10 pts, c) 15 pts.

a) Given Fig. 1 and $P(w_1) = 0.2$, what would be the Bayes decision rule?

b) Given Fig. 1, $P(w_1) = 0.5$ and the following loss functions: $\lambda(\alpha_1|w_2) = 0.5, \lambda(\alpha_2|w_1) = 1, \lambda(\alpha_i|w_j) = 0, \forall i = j$, what would be the Bayes decision rule that minimizes the overall risk?
(Hint: find the new decision boundary first.)

c) Compute the Bayes risk with the decision rule given in b).

Q3. (20 pts) Maximum Likelihood Estimation.

Given n random variables $x_1, \ldots, x_n$, each of which is obtained by $x_i = \beta z_i + \epsilon_i, i \in [1, n]$, where $z_1, \ldots, z_n$ are known fixed constants, and $\forall i, \epsilon_i \sim N(0, \sigma^2)$. σ^2 is known. Please give the maximum likelihood estimation of β. (Hint: $x_i \sim N(\beta z_i, \sigma^2)$.)

Q4. (25 pts) Maximum Likelihood and Maximum A Posteriori: a) 10 pts, b) 15pts.

Maximum likelihood methods could also be used to estimate the class prior probabilities. Let samples are *i.i.d.* drawn from a class w_j with unknown probability $P(w_j) = \theta_j, 0 \leq \theta_j \leq 1$. Let $z_{ij} = 1$ if the class for the i-th sample is w_j and $z_{ij} = 0$ otherwise. We have the likelihood function as

$$P(z_{1j}, \ldots, z_{nj}|\theta_j) = \prod_{i=1}^{n} \theta_j^{z_{ij}} (1 - \theta_j)^{1 - z_{ij}},$$

where n is the number of samples.

a) What's the maximum likelihood estimation for θ_j?

b) Let $\Theta = [\theta_1, \ldots, \theta_j, \ldots, \theta_K]^{\mathrm{T}} \in \mathbb{R}^K$ be the probability distribution over K class. We assume that the prior of Θ is a Dirichlet distribution, *i.e.*, $p(\Theta) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$, where $\alpha_k, 1 \leq k \leq K$, are K fixed hyperparameters and $B(\alpha)$ is a constant w.r.t. θ. What's the MAP estimation for θ_j?
(Hint: $\frac{\partial}{\partial \theta_j} \log p(\theta_j) = \frac{\partial}{\partial \theta_j} \log p(\Theta)$)

1.

(1) AD (2) BD (3) AC (4) AC (5) BC

(6) Ridge regression solution:

$$L(w) = \frac{1}{2}\sum_{n=1}^{N}(y_n - w^T x_n)^2 + \frac{\lambda}{2}\|w\|_2^2$$

$$\|w\|_2^2 = w^T w$$

$$L(w) = \frac{1}{2}\left(\sum_{n=1}^{N}(y_n - w^T x_n)^2 + \lambda w^T w\right)$$

$$\frac{2}{2w}L(w) = \frac{1}{2}\left(-2\sum_{n=1}^{N}(y_n - w^T x_n)x_n + 2\lambda w\right)$$

$$= \frac{1}{2}\left(-2\sum_{n=1}^{N} y_n x_n + 2\sum_{n=1}^{N}(w^T x_n)x_n + 2\lambda w\right)$$

$$= -X^T y + X^T X w + \lambda w$$

$$\min_w L(w) \Rightarrow \frac{2}{2w}L(w) = 0$$

$$\Rightarrow -X^T y + X^T X w + \lambda w = 0$$

$$w = (X^T X + \lambda I)^{-1} X^T y$$

GD can lead to a global optima for Ridge regression
Because it's a convex optimization which GD can guarantee global optima for convex.

(7)

MLE = Maximum Likelihood Estimation, which point estimation by maximizing the likelihood $P(D|\theta)$ $\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$. It starts with the probability of observation given the parameter. It don't take consideration of the prior knowledge.

MAP = Maximum a posterior. which point estimation by maximizing the probability $P(D|\theta)P(\theta)$. MAP can take into account prior knowledge about what we expect our parameters to be in the form of a prior probability distribution.

Full Bayesian parameter estimation will take consider all the θ. which is not a point estimation.
treat θ as a random variable with prior $P(\theta)$, compute the full posterior distribution $P(\theta|D)$

2.

a.

posterior $P(w_1|x) = P(w_2|x)$

$$\Downarrow$$

$$P(x|w_1) \cdot P(w_1) = P(x|w_2) P(w_2)$$

we know $P(w_1) = 0.2$. Thus $P(w_2) = 1 - 0.2 = 0.8$

$$(-2x+2) \cdot 0.2 = (x - \tfrac{1}{2}) \cdot 0.8$$

$$-\frac{2}{5}x + \frac{2}{5} = \frac{4}{5}x - \frac{2}{5}$$

$$\frac{6}{5}x = \frac{4}{5}$$

$$x = \frac{2}{3}$$

Choose w_1 if $x < \frac{2}{3}$

choose w_2 if $x > \frac{2}{3}$

b.

Conditional risk formula for two category classification:

$$R(\alpha_1|x) = \lambda_{11}P(w_1|x) + \lambda_{12}P(w_2|x)$$

$$R(\alpha_2|x) = \lambda_{21}P(w_1|x) + \lambda_{22}P(w_2|x)$$

The prior of two classes are equal

$$P(w_1) = P(w_2) = \frac{1}{2}$$

$\because \lambda(\alpha_i|w_j) = 0, \forall i = j$

$\therefore \lambda_{11} = 0 \quad \lambda_{22} = 0$ (choosing correctly)

$\lambda_{12} = 0.5 \quad \lambda_{21} = 1$ (choosing correctly

Thus the expression of conditional risk is:

$$R(\alpha_1|x) = \lambda_{11}P(w_1|x) + \lambda_{12}P(w_2|x)$$

$$= 0\,P(w_1|x) + 0.5\,P(w_2|x)$$

$$= 0.5\,P(w_2|x)$$

$$R(\alpha_2|x) = \lambda_{21} P(w_1|x) + \lambda_{22} P(w_2|x)$$

$$= 1 \cdot P(w_1|x) + 0 \cdot P(w_2|x)$$

$$= P(w_1|x)$$

Let $R(\alpha_1|x) = R(\alpha_2|x)$

$$0.5 P(w_2|x) = P(w_1|x)$$

$$\frac{0.5 P(x|w_2) P(w_2)}{P(x)} = \frac{P(x|w_1) P(w_1)}{P(x)}$$

$\therefore$ priors $P(w_1) = P(w_2) = 0.5$

$\therefore$ $0.5 P(x|w_2) = P(x|w_1)$

$$0.5\left(x - \frac{1}{2}\right) = -2x + 2$$

$$\frac{1}{2}x - \frac{1}{4} = -2x + 2$$

$$\frac{5}{2}x = \frac{9}{4}$$

$$x = \frac{9}{10}$$

C. The decision region for R_1 is $x < \frac{9}{10}$

decision region for R_2 is $x > \frac{9}{10}$

$$Risk = \int_{R_1} \lambda_{11} P(w_1|x) + \lambda_{12} P(w_2|x) + \int_{R_2} \lambda_{21} P(w_1|x) + \lambda_{22} P(w_2|x)$$

$$= \int_{R_1} \lambda_{12} P(x|w_2) P(w_2) + \int_{R_2} \lambda_{21} P(x|w_1) P(w_1)$$

$$= \int_{x=0.5}^{x=\frac{9}{10}} (0.5) \cdot P(x|w_2) \cdot (0.5) + \int_{x=\frac{9}{10}}^{x=1} 1 \cdot P(x|w_1) \cdot (0.5)$$

$$= (0.25) \cdot 0.08 + (0.5) \cdot 0.01$$

$$= 0.02 + 0.005$$

$$= 0.025$$

3.

Because $X_i = \beta z_i + \varepsilon_i \quad i \in [1, n] \quad z_1 \cdots z_n$ are fixed constants.

If $z = z$, then $X = \beta z + \epsilon$ for some parameter β and some random noise variable ϵ.

$$\epsilon \sim N(0, \sigma^2)$$

Then the model tells us the conditional pdf of X for each z.

$$p(X \mid z = z; \beta, \sigma^2)$$

Given any data set $(z_1, x_1)(z_2, x_2) \cdots (z_n, x_n)$

we can write down the probability density:

$$\prod_{i=1}^{n} p(x_i \mid z_i; \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \beta z_i)^2}{2\sigma^2}}$$

log-likelihood:

$$l(\beta, \sigma^2) = \ln \prod_{i=1}^{n} p(x_i \mid z_i; \beta, \sigma^2)$$

$$= \sum_{i=1}^{n} \ln p(x_i \mid z_i; \beta, \sigma^2)$$

$$= -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \beta z_i)^2}{2\sigma^2}$$

Take partial derivative of β to 0 to find the MLE:

$$\frac{\partial}{\partial \beta}\left(-\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \beta z_i)^2}{2\sigma^2}\right)$$

$$= \sum_{i=1}^{n} -\frac{z_i(x_i - \beta z_i)}{\sigma^2} = 0$$

$$\beta = \sum_{i=1}^{n}\left(\frac{z_i x_i}{z_i^2}\right)$$

4.

a)

log likelihood function of θ_j is:

$$l(\theta_j) = \ln P(z_{1j}, \dots z_{nj} \mid \theta_j)$$

$$= \ln\left(\prod_{i=1}^{n} \theta_j^{z_{ij}} (1-\theta_j)^{1-z_{ij}}\right)$$

$$= \sum_{i=1}^{n} \left(z_{ij} \ln(\theta_j) + (1-z_{ij}) \ln(1-\theta_j)\right)$$

take the partial gradient of θ_j we get

$$\nabla_{\theta_j} l(\theta_j) = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \left(z_{ij} \ln(\theta_j) + (1-z_{ij}) \ln(1-\theta_j)\right)$$

$$= \frac{1}{\theta_j} \sum_{i=1}^{n} z_{ij} - \frac{1}{1-\theta_j} \sum_{i=1}^{n} (1-z_{ij})$$

$$= 0$$

solve the equation we got:

$$(1-\theta_j) \sum_{i=1}^{n} z_{ij} = \theta_j \sum_{i=1}^{n} (1-z_{ij})$$

$$\Downarrow$$

$$\sum_{i=1}^{n} z_{ij} = \theta_j \sum_{i=1}^{n} z_{ij} + n\theta_j - \theta_j \sum_{i=1}^{n} z_{ij}$$

$$= n\theta_j$$

Thus $\theta_j = \frac{1}{n} \sum_{i=1}^{n} z_{ij}$

b)

We have the MAP estination as:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|D)$$

$$= \arg\max_{\theta} \log P(D|\theta) + \log p(\theta)$$

log likelihood function of θ_j is:

$$l(\theta_j) = \ln P(z_{1j}, \dots z_{nj} \mid \theta_j) + \ln P(\theta_j)$$

$$= \ln\left(\prod_{i=1}^{n} \theta_j^{z_{ij}} (1-\theta_j)^{1-z_{ij}}\right) + \ln p(\theta_j)$$

$$= \sum_{i=1}^{n} \left(z_{ij} \ln(\theta_j) + (1-z_{ij}) \ln(1-\theta_j)\right) + \ln P(\theta_j)$$

$$\ln P(\theta) = \ln\left(\frac{1}{B(a)} \prod_{k=1}^{k} \theta_k^{a_k-1}\right)$$

$$= \sum_{k=1}^{K} \ln\left(\frac{1}{B(a)} \theta_k^{a_k-1}\right)$$

take the gradient we got

$$\frac{\partial}{\partial \theta_j} \ln p(\theta) = \sum_{i=1}^{n} \frac{a_i - 2}{\theta_j}$$

Because $\frac{\partial}{\partial \theta_j} \log p(\theta_j) = \frac{\partial}{\partial \theta_j} \log p(\theta)$

Thus

$$\nabla_{\theta_j} L(\theta_j) = \frac{1}{\theta_j} \sum_{i=1}^{n} z_{ij} - \frac{1}{1-\theta_j} \sum_{i=1}^{n} (1 - z_{ij}) + \sum_{i=1}^{n} \frac{\alpha_i - 1}{\theta_j}$$

$$= 0$$

Solve this equation we got

$$\frac{1}{\theta_j} \sum_{i=1}^{n} z_{ij} + \sum_{i=1}^{n} \frac{\alpha_i - 1}{\theta_j} = \frac{1}{1-\theta_j} \sum_{i=1}^{n} (1 - z_{ij})$$

$$\frac{1-\theta_j}{\theta_j} \sum_{i=1}^{n} z_{ij} + (1-\theta_j) \sum_{i=1}^{n} \frac{\alpha_i - 1}{\theta_j} = n - \sum_{i=1}^{n} z_{ij}$$

$$\frac{1-\theta_j}{\theta_j} \sum_{i=1}^{n} z_{ij} + \frac{1-\theta_j}{\theta_j} \sum_{i=1}^{n} \frac{\alpha_i - 1}{\theta_j} = n - \sum_{i=1}^{n} z_{ij}$$

$$\frac{1-\theta_j}{\theta_j} \left(\sum_{i=1}^{n} z_{ij} + \alpha_i - 1 \right) = n - \sum_{i=1}^{n} z_{ij}$$

$$\frac{1-\theta_j}{\theta_j} = \frac{n - \sum_{i=1}^{n} z_{ij}}{\sum_{i=1}^{n} z_{ij} + \alpha_i - 1} \Rightarrow \theta_j = \frac{1}{1 + \frac{n - \sum_{i=1}^{n} z_{ij}}{\sum_{i=1}^{n} z_{ij} + \alpha_i - 1}}$$