AI 534: Machine Learning

Assignment 1

Name : Nam Nguyen

ID: 934-422-327

Problem 1: $X_1, X_2, \ldots, X_n \sim U(0, \theta)$

a) The likelihood function of θ:

$$L = P(X_1, \ldots, X_n \mid \theta) = \prod_{i=1}^{n} p(X_i \mid \theta)$$

$$= \prod_{i=1}^{n} \frac{1}{\theta} I_{\{X_i \leq \theta\}} = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{if } \forall X_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

where I_A is the indicator function of event A

b) Derive the maximum likelihood estimation for θ:

$$\arg\max_{\theta} L = \arg\max_{\theta} P(X_1, \ldots, X_n \mid \theta)$$

$$= \arg\max_{\theta} \left(\frac{1}{\theta}\right)^n \quad \text{where } \forall X_i \leq \theta$$

$$= \max\{X_1, \ldots, X_n\}$$

Therefore, $\theta^* = \max\{X_1, \ldots, X_n\}$

Problem 2:

a) The likelihood function for the linear regression model can be written as

$$L(\underline{w}; \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(- \frac{(y_i - \underline{w}^T x_i)^2}{2\sigma_i^2}\right)$$

We get the log-likelihood by taking the logarithm of the likehood function

$$\log L(\underline{w}; \sigma^2) = -\frac{1}{2} \sum_{i=1}^{n} \left[\log(2\pi\sigma_i^2) + \frac{(y_i - \underline{w}^T w_i)^2}{\sigma_i^2}\right] \quad (1)$$

b)

- We have, weighted square loss function

$$J(\underline{w}) = \sum_{i=1}^{n} a_i (\underline{w}^T x_i - y_i)^2$$

- From (1) we have,

$$\log L(\underline{w}; \sigma^2) = \underbrace{-\frac{1}{2} \sum_{i=1}^{n} \log(2\pi\sigma_i^2)}_{\text{constant}} - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \underline{w}^T w_i)^2}{\sigma_i^2}$$

Therefore, maximize the log-likelihood is equivalent to minimie $\sum_{i=1}^{n} \frac{1}{\delta_i^2} (y_i - \underline{w}^T w_i)^2$

Hence, $\boxed{a_i = \frac{1}{\delta_i^2}}$

c) Derive the batch gradient descent update rule for optimize this objective.

We have,

$$\nabla_{\underline{w}} J(\underline{w}) = 2 \sum_{i=1}^{n} a_i (\underline{w}^T x_i - y_i) x_i$$

We update $\underline{w}$ as follows

$$\underline{w}^{(t+1)} = \underline{w}^{(t)} - \alpha \nabla_{\underline{w}} J(\underline{w})$$

$$= \underline{w}^{(t)} - \alpha . \left(2 \sum_{i=1}^{n} a_i (\underline{w}^T x_i - y_i) x_i \right)$$

where α is learning rate and t denotes the iteration step.

d) Derive the close form solution to this optimization problem:

To find a close-form solution, we can rewrite the weighted square loss function $J(\underline{w})$ in matrix form using a diagonal matrix A, with $A(i, i) = \alpha_i \cdot a_i$.

The new loss function is

$$J(\underline{W}) = (\underset{\sim}{X}\underline{W} - \underline{y})^T \underset{\sim}{A} (\underset{\sim}{X}\underline{W} - \underline{Y}) \quad (*)$$

where $\underset{\sim}{X} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1d} \\ \vdots & & & \\ x_{N0} & x_{N1} & \cdots & x_{Nd} \end{bmatrix}$; $\underline{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}$

$$\underset{\sim}{A} = \begin{bmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_N \end{bmatrix}$$

We can see that $(*)$ is a weighted Least Square problem. The close-form solution for $(*)$ can be expressed as follows (by solving $\nabla J(\underline{W}) = \underline{0}$)

$$\hat{\underline{W}}_{WLS} = (\underset{\sim}{X}^T \underset{\sim}{A} \underset{\sim}{X})^{-1} \underset{\sim}{X}^T \underset{\sim}{A} \underline{y}.$$

Problem 3:

a) We have,

Expected cost = p (classified as spam). Cost (spam → spam)

+ p (classified as non-spam). Cost (non-spam → spam)

$= (0.8)(0) + (1-0.8)(10) = 2.$

b)

- We should classify the email base on the threshold that minimizes the expected mis-classification cost.
- If the predicted probability, $p \geq \theta$ then

  classify the email as spam

  else

  classify the email as non-spam.

c)

- To minime the expected mis-classification cost, we need to choose a threshold (θ) such that the expected cost is minimized.
- Calculate the expected cost for different values of threshold θ ⟹ Choose the θ_{min} that minimize it.

- If the predicted probability $p \geq \theta$ then
  classify the email as spam
  Else
  classify the email as non-spam

d)

- The cost matrix.

predicted label $\hat{y}$	True label y	
	non-spam	spam
non-spam	0	1
spam	5	0

- Hence,

The cost of classifying a non-spam as spam is 5. (1)

The cost of classifying a spam as non-spam is 1. (2)

- If $p < 1/5$, it's more cost-effective to classify as non-spam to avoid the high error cost (1)

- If $p \geq 1/5$, it's more cost-effective to classify as spam to avoid the low error cost (2).

This cost matrix gives $\theta = 1/5$.

Problem 4 :

a) The posterior distribution is written by

$$p(\hat{\theta} \mid X_1, \ldots, X_n, \alpha, \beta) = \frac{P(X_1, \ldots, X_n \mid \theta = \hat{\theta})\, p(\theta = \hat{\theta} \mid \alpha, \beta)}{P(X_1, \ldots, X_n)}$$

- We have,

$$P(X_1, \ldots, X_n \mid \theta = \hat{\theta}) = \hat{\theta}^{k} (1 - \hat{\theta})^{n-k}$$

where $\begin{cases} k \text{ is the number of observed successes} \\ n \text{ is the total number of observations} \end{cases}$

- And,

$$p(\theta = \hat{\theta} \mid \alpha, \beta) = \frac{\hat{\theta}^{(\alpha-1)} (1 - \hat{\theta})^{(\beta-1)}}{B(\alpha, \beta)}$$

Hence,

$$p(\hat{\theta} \mid X_1, \ldots, X_n, \alpha, \beta) = \frac{\hat{\theta}^{k} (1 - \hat{\theta})^{n-k} \cdot \hat{\theta}^{(\alpha-1)} (1 - \hat{\theta})^{(\beta-1)}}{P(X_1, \ldots, X_n)\, B(\alpha, \beta)}$$

$$= \frac{\hat{\theta}^{(k+\alpha-1)} (1 - \hat{\theta})^{n-k+\beta-1}}{P(X_1, \ldots, X_n)\, B(\alpha, \beta)}$$

Therefore, it is also the probability density function of a Beta distribution with $\alpha' = k + \alpha$ and $\beta' = n - k + \beta$.

b)

- Observing 5 coin tosses with 2 of them being heads.

The posterior distribution of θ is

$$Beta(2+2, 5-2+2) = Beta(4,5)$$

- Observing 50 coin tosses with 20 of them being heads

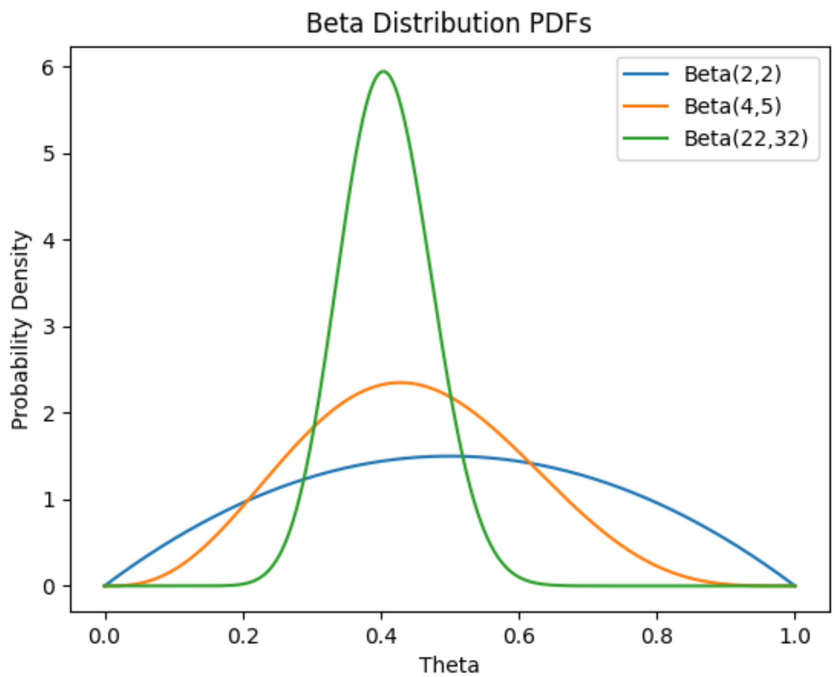
The posterior distribution of θ is

$$Beta(2+50, 50-20+2) = Beta(22, 32)$$

- Plot pdf functions:

- Expectation:
  + Prior distribution, $Beta(2,2)$: A symmetric distribution with a peak around 0.5
  + Posterior distribution, $Beta(4,5)$: A shifted and slightly narrower distribution compared to the prior.
  + Posterior distribution, $Beta(22, 32)$: A narrower and more peaked distribution that is centered around the true value, $\theta = 0.4$.

- Conclusion: As we observe more coin tosses, the posterior becomes more concentrates around the true value of $\theta = 0.4$ since the more data reduces the impact of the prior



Beta Distribution PDFs
Beta(2,2)
Beta(4,5)
Beta(22,32)
Probability Density
Theta
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0

Problem 5 :

a) Case 1 : $\underline{w}_0 = \underline{0}$

The Perception algorithmm will misclassify this point $\underline{x}$ only once before convergence.

It will update $\underline{w}$ in the correct direction after the first misclassification.

b) Case 2 : $\underline{w}_0 \; != 0$

The number of times the Perception algorithm mis ~~classifyis~~ mis-classifies this point $\underline{x}$ before convergence depends on the angle between the initial weight vector, $\underline{w}_0$ and the data point ~~Xd~~ $\underline{x}$.

$$\Rightarrow \quad \cos\theta = \frac{\underline{w}_0 \cdot \underline{x}}{\|\underline{w}_0\|_2 \; \|\underline{x}\|_2}.$$

$$\text{Number of misclassifications} = \frac{\theta}{\cos(\theta)}.$$

Problem 6:

- The likelihood function:

$$L(\underline{w}) = \prod_{i=1}^{N} \prod_{k=1}^{K} p(y=k \mid \underline{x}_i)^{y_{ik}}$$

- The log-likelihood function:

$$\log L(\underline{w}) = \log \prod_{i=1}^{N} \prod_{k=1}^{K} p(y=k \mid \underline{x}_i)^{y_{ik}}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \log p(y=k \mid \underline{x}_i)$$

- We have,

$$\log p(y=k \mid \underline{x}_i) = \log \left[\frac{\exp(\underline{w}_k^T \underline{x}_i)}{\sum_{j=1}^{K} \exp(\underline{w}_j^T \underline{x}_i)} \right]$$

$$= \log \left[\exp(\underline{w}_k^T \underline{x}_i) \right] - \log \left[\sum_{j=1}^{K} \exp(\underline{w}_j^T \underline{x}_i) \right]$$

- Hence,

$$\log L(\underline{w}) = \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \left[\log\left(\exp(\underline{w}_k^T \underline{x}_i)\right) - \log\left(\sum_{j=1}^{K} \exp(\underline{w}_j^T \underline{x}_i) \right) \right]$$

- Let $z_i = \sum_{j=1}^{K} \exp(\underline{w}_j^T \underline{x}_i)$

- The gradient of the log-likelihood

$$\nabla_{\underline{w}_c} \log L(\underline{w}) = \sum_{i=1}^{N} y_{ic} \underline{x}_i - \frac{\exp(\underline{w}_c^T \underline{x}_i)}{z_i} \sum_{k=1}^{K} y_{ik} \underline{x}_i$$

$$= \sum_{i=1}^{N} y_{ic} \underline{x}_i - \frac{\exp(\underline{w}_c^T \underline{x}_i)}{z_i} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \underline{x}_i .$$