ENGR 421/DASC 521: Introduction to Machine Learning
Spring 2021 Final – Solution Key

Question 1:
What are the differences between parametric methods and non-parametric methods? Name some of the parametric and non-parametric machine learning algorithms.

- **Parametric**: Number of parameters are not affected by the training data set size, stronger assumptions about the data
  Examples: linear/logistic regression, linear discriminant analysis, perceptron, neural networks
- **Non-parametric**: Number of parameters grows with the size of the training set, fewer assumptions about the data
  Examples: k-nearest neighbours, decision trees, support vector machines

Question 2:
For a split used for a numeric input in a decision tree, instead of a binary split, how can we increase the number of children to three? What are the advantages and the disadvantages of such a split over a binary split?

One can use a ternary split with two thresholds and three branches as

$$x_j < w_{ma}, \quad w_{ma} \leq x_j < w_{mb}, \quad x_j \geq w_{mb}.$$

- One ternary node splits an input into three, whereas this requires two successive binary nodes.
- The complexity of finding the best pair grows larger and each node stores two thresholds instead of one and has three branches instead of two.

Question 3:
In regression problems, we calculate the error as the sum of squared differences between the true and predicted values. This particular choice is the most frequently used error function, but it is not robust to outliers since it increases rapidly with the increasing size of the difference.

(a) What would be a better error function to implement robust regression?

(b) Why do we usually prefer working with the sum of squared differences?

(a) One alternative is to use the absolute value (or ϵ-insensitive loss) as an error function instead of squaring the differences.

(b) This achieves robustness, but is hard to work with in practice because the absolute value function is not differentiable.

Question 4:
Suppose that you are given the source code of a decision tree algorithm for classification. Describe how you can modify the source code to obtain a decision tree algorithm for regression.

First, we should find an impurity measure suitable for regression problems such as squared error. We should then find a suitable stopping condition for determining leaf nodes.

Question 5:
Explain the effect of the dimensionality of the reduced space on the reconstruction error in principal component analysis (PCA).

As the dimensionality of the reduced space increases, the reconstruction error decreases. However, this decrease is very fast at first and slows later on.

Question 6:
What is the difference between principal component analysis (PCA) and linear discriminant analysis (LDA)?

Principal Component Analysis: It is an unsupervised learning method that aims to find the directions that maximize the variance of the data.

Linear Discriminant Analysis: It is a supervised learning method that aims to find the directions that best separate the classes.

Question 7:
Clustering is an unsupervised and iterative algorithm that has two steps:

(1) Estimate the labels,

(2) Estimate the parameters.

k-means and Expectation–Maximization (EM) are special cases of clustering algorithm. What are the labels and the parameters for these algorithms in (1) and (2), respectively.

For k-means, labels are cluster memberships (it is a binary variable), and parameters are cluster means (i.e., sample mean of each cluster).

For EM, labels are cluster memberships (it is a random variable), and parameters are cluster means and covariances (i.e., sample mean and covariances of each cluster).

Question 8:
Assume that we have two classes that are exponentially distributed with rates λ_1 and λ_2, and prior probabilities $P(y = 1)$ and $P(y = 2)$. The probability density function of exponential distribution with rate λ is $p(x) = \lambda e^{-\lambda x}$. Determine the parameters of the following discriminant function.

$$g(x) = \log \frac{P(y = 1|x)}{P(y = 2|x)} = w_1 x + w_0$$

$$\begin{aligned} g(x) &= \log \left[\frac{P(y = 1|x)}{P(y = 2|x)} \right] \\ &= \log \left[\frac{p(x|y = 1)P(y = 1)}{p(x|y = 2)P(y = 2)} \right] \end{aligned}$$

$$
\begin{aligned}
&= \log\left[\frac{\lambda_1 e^{-\lambda_1 x} P(y=1)}{\lambda_2 e^{-\lambda_2 x} P(y=2)}\right] \\
&= \log\left[e^{(\lambda_2-\lambda_1)x}\right] + \log\left[\frac{\lambda_1 P(y=1)}{\lambda_2 P(y=2)}\right] \\
&= \underbrace{(\lambda_2-\lambda_1)}_{w_1} x + \underbrace{\log\left[\frac{\lambda_1 P(y=1)}{\lambda_2 P(y=2)}\right]}_{w_0}
\end{aligned}
$$

$$
\begin{aligned}
w_1 &= \lambda_2 - \lambda_1 \\
w_0 &= \log\left[\frac{\lambda_1 P(y=1)}{\lambda_2 P(y=2)}\right]
\end{aligned}
$$

Question 9:
Let $\mathcal{X} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ be the training set that contains N data points for a binary classification problem. The primal optimization problem for support vector machine can be written as follows:

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^N \gamma_i \xi_i \\
\text{with respect to} \quad & \boldsymbol{w},\ \boldsymbol{\xi},\ b \\
\text{subject to} \quad & y_i\big(\boldsymbol{w}^\top \boldsymbol{x}_i + b\big) \geq 1 - \xi_i \quad \forall i \\
& \xi_i \geq 0 \quad \forall i
\end{aligned}
$$

where $\boldsymbol{\xi}$ is the vector of slack variables, C is the regularization parameter, and γ_i is a given coefficient for balancing the effect of errors for different classes based on the class populations. Derive the dual optimization problem.

$$
\mathcal{L}_P = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^N \gamma_i \xi_i - \sum_{i=1}^N \alpha_i\big(y_i\big(\boldsymbol{w}^\top \boldsymbol{x}_i + b\big) - 1 + \xi_i\big) - \sum_{i=1}^N \beta_i \xi_i
$$

$$
\frac{\partial \mathcal{L}_P}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^N \alpha_i y_i \boldsymbol{x}_i = 0 \qquad \Rightarrow \qquad \boldsymbol{w} = \sum_{i=1}^N \alpha_i y_i \boldsymbol{x}_i
$$

$$\frac{\partial \mathcal{L}_P}{\partial b} = -\sum_{i=1}^{N} \alpha_i y_i = 0 \qquad \Rightarrow \qquad \sum_{i=1}^{N} \alpha_i y_i = 0$$
$$\frac{\partial \mathcal{L}_P}{\partial \xi_i} = C\gamma_i - \alpha_i - \beta_i = 0 \quad \forall i \qquad \Rightarrow \qquad 0 \leq \alpha_i \leq C\gamma_i \quad \forall i$$

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j \\ \text{with respect to} \quad & \boldsymbol{\alpha} \\ \text{subject to} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C\gamma_i \quad \forall i \end{aligned}$$

Question 10:
Assume that we have two classes from two multivariate Gaussian distributions with different mean vectors and equal covariance matrices; that is, we have $p(\boldsymbol{x}|y=1) \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $p(\boldsymbol{x}|y=2) \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Assume also that the prior probabilities are equal; that is, we have $P(y=1) = P(y=2)$. Derive the position of the intersection of the two posterior probabilities.

$$\begin{aligned} P(y=1|\boldsymbol{x}) &= P(y=2|\boldsymbol{x}) \\ p(\boldsymbol{x}|y=1)\underbrace{P(y=1)}_{1/2} &= p(\boldsymbol{x}|y=2)\underbrace{P(y=2)}_{1/2} \\ \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right] &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right] \\ (\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) &= (\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \\ (\boldsymbol{x}-\boldsymbol{\mu}_1) &= -(\boldsymbol{x}-\boldsymbol{\mu}_2) \\ \boldsymbol{x}^\star &= \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \end{aligned}$$

Question 11:
Let $\mathcal{X} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ be the training set that contains N multivariate data points for a linear regression problem. Let $\mathcal{L}(\boldsymbol{w}, b)$ be the regularized loss function:

$$\mathcal{L}(\boldsymbol{w}, b) = \frac{1}{N}\sum_{i=1}^{N} l(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i - b) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$$

where λ is a regularization constant, and $l(\cdot)$ is defined as

$$l(\xi) = \begin{cases} \frac{1}{2}\xi^2 & \text{if } |\xi| < 1 \\ |\xi| - \frac{1}{2} & \text{otherwise.} \end{cases}$$

Calculate the gradient of the regularized loss function with respect to $\boldsymbol{w}$.

$$\frac{\partial l(\xi)}{\partial \xi} = \begin{cases} \xi & \text{if } |\xi| < 1 \\ \text{sign}(\xi) & \text{otherwise} \end{cases}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{w}, b)}{\partial \boldsymbol{w}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial l(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i - b)}{\partial \boldsymbol{w}} + \lambda \boldsymbol{w}$$

$$\frac{\partial l(y_i - \boldsymbol{w}^\top \boldsymbol{x} - b)}{\partial \boldsymbol{w}} = \begin{cases} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i - b)(-\boldsymbol{x}_i) & \text{if } |y_i - \boldsymbol{w}^\top \boldsymbol{x}_i - b| < 1 \\ \text{sign}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i - b)(-\boldsymbol{x}_i) & \text{otherwise} \end{cases}$$

Question 12:
Consider the following training data set for a binary classification problem:

positively labeled data points: $(3, 1)$, $(3, -1)$, $(6, 1)$, $(6, -1)$

negatively labeled data points: $(1, 0)$, $(0, 1)$, $(0, -1)$, $(-1, 0)$

Write the primal and dual support vector machine problems corresponding to this problem instance. Give the optimal solutions to the primal and dual optimization problems. Identify the support vectors. Write the separating hyperplane in terms of the support vectors.

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2}(w_1^2 + w_2^2) \\
\text{with respect to} \quad & w_1,\ w_2,\ b \\
\text{subject to} \quad & +1(+3w_1 + 1w_2 + b) \geq 1 \\
& +1(+3w_1 - 1w_2 + b) \geq 1 \\
& +1(+6w_1 + 1w_2 + b) \geq 1 \\
& +1(+6w_1 - 1w_2 + b) \geq 1 \\
& -1(+1w_1 + 0w_2 + b) \geq 1 \\
& -1(+0w_1 + 1w_2 + b) \geq 1 \\
& -1(+0w_1 - 1w_2 + b) \geq 1 \\
& -1(-1w_1 + 0w_2 + b) \geq 1
\end{aligned}
$$

$$w_1^\star = 1,\ w_2^\star = 0,\ b^\star = -2$$

$$
\begin{aligned}
\text{maximize} \quad & \sum_{i=1}^{8} \alpha_i - \frac{1}{2}\begin{bmatrix}\alpha_1 & \alpha_2 & \dots & \alpha_8\end{bmatrix}\begin{bmatrix} y_1y_1\boldsymbol{x}_1^\top\boldsymbol{x}_1 & y_1y_2\boldsymbol{x}_1^\top\boldsymbol{x}_2 & \dots & y_1y_8\boldsymbol{x}_1^\top\boldsymbol{x}_8 \\ y_2y_1\boldsymbol{x}_2^\top\boldsymbol{x}_1 & y_2y_2\boldsymbol{x}_2^\top\boldsymbol{x}_2 & \dots & y_2y_8\boldsymbol{x}_2^\top\boldsymbol{x}_8 \\ \vdots & \vdots & \ddots & \vdots \\ y_8y_1\boldsymbol{x}_8^\top\boldsymbol{x}_1 & y_8y_2\boldsymbol{x}_8^\top\boldsymbol{x}_2 & \dots & y_8y_8\boldsymbol{x}_8^\top\boldsymbol{x}_8 \end{bmatrix}\begin{bmatrix}\alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_8\end{bmatrix} \\
\text{with respect to} \quad & \alpha_1,\ \alpha_2,\ \alpha_3,\ \alpha_4,\ \alpha_5,\ \alpha_6,\ \alpha_7,\ \alpha_8 \\
\text{subject to} \quad & +\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \alpha_5 - \alpha_6 - \alpha_7 - \alpha_8 = 0 \\
& \alpha_i \geq 0 \quad i = 1, 2, \dots, 8
\end{aligned}
$$

$$\alpha_1^\star = 1/4,\ \alpha_2^\star = 1/4,\ \alpha_3^\star = 0,\ \alpha_4^\star = 0,\ \alpha_5^\star = 1/2,\ \alpha_6^\star = 0,\ \alpha_7^\star = 0,\ \alpha_8^\star = 0$$

Question 13:
Let $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^N$ be the training set that contains N univariate data points for a regression problem, which is described by the following model:

$$y_i = ax_i + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(\epsilon_i; 0, \sigma^2)$, and each data point is independent of the others.

(a) Write the likelihood function $p(y_1, y_2, \dots, y_N | x_1, x_2, \dots, x_N, a, \sigma^2)$.

(b) Compute the maximum likelihood estimate of a.

$$\epsilon_i \sim \mathcal{N}(\epsilon_i; 0, \sigma^2) \Rightarrow y_i \sim \mathcal{N}(y_i; ax_i, \sigma^2)$$

(a)

$$p(y_1, y_2, \ldots, y_N | x_1, x_2, \ldots, x_N, a, \sigma^2) = \prod_{i=1}^{N} p(y_i | x_i, a, \sigma^2)$$
$$= \prod_{i=1}^{N} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - a x_i)^2}{2\sigma^2} \right) \right]$$

(b)

$$\text{log-likelihood} = \sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - a x_i)^2}{2\sigma^2} \right) \right]$$
$$=^{+} \sum_{i=1}^{N} -\frac{(y_i - a x_i)^2}{2\sigma^2}$$

$$\frac{\partial \text{log-likelihood}}{\partial a} = \sum_{i=1}^{N} -\frac{2(y_i - a x_i)(-x_i)}{2\sigma^2} = 0 \Rightarrow a^\star = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$$