Homework Set 2

Introduction to Artificial Intelligence with Mathematics (MAS473)

Total Points = 50pts

1. (15pts) **(Bayesian Linear Regression)** Consider a linear model

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^\top \phi(\mathbf{x}) + \epsilon$$

where $\phi(\mathbf{x}) = [1, \mathbf{x}^\top]^\top \in \mathbb{R}^{D+1}$, $\mathbf{w} \in \mathbb{R}^{D+1}$ and $\epsilon \sim \mathcal{N}(0, \beta^{-1})$ is an independent Gaussian noise term. ($\beta > 0$ is a given hyperparameter) Let $\mathcal{D} = \{(\mathbf{x}_i, y_i) : 1 \leq i \leq N\}$ be a given dataset,

$$\Phi = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \cdots, \phi(\mathbf{x}_N)]^\top \in \mathbb{R}^{N \times (D+1)}$$

(it is called a design matrix) and $\mathbf{y} = [y_1, \cdots, y_N]^\top \in \mathbb{R}^N$. Note that ϕ could be an arbitrary non-linear mapping in general.

(a) Suppose $\mathbf{w}$ has a prior distribution $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \alpha^{-1} I)$ where $\alpha > 0$ is a given hyperparameter. Find the posterior distribution $p(\mathbf{w}|\mathcal{D})$ and the maximum a posterior (MAP) estimator of $\mathbf{w}$. Note that finding the MAP estimator is equivalent to using the Ridge regression in this case.

(b) Prove that the MAP estimator of $\mathbf{w}$ converges to the MLE estimator of $\mathbf{w}$ as $\alpha \to 0$, i.e. $\mathbf{w}_{MAP} \to \mathbf{w}_{MLE}$ as $\alpha \to 0$.

(c) Note that the prediction distribution at point $\mathbf{x}_0$ can be calculated by

$$p(y_0|\mathcal{D}, \mathbf{x}_0) = \int p(y_0|\mathbf{w}, \mathbf{x}_0) p(\mathbf{w}|\mathcal{D}) \, d\mathbf{w}.$$

Prove that

$$p(y_0|\mathcal{D}, \mathbf{x}_0) = \mathcal{N}(y_0|\beta \cdot \phi(\mathbf{x}_0)^\top (\alpha I + \beta \Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}, 1/\beta + \phi(\mathbf{x}_0)^\top (\alpha I + \beta \Phi^\top \Phi)^{-1} \phi(\mathbf{x}_0)).$$

Solution.

(a) Note that $p(\mathbf{w}|\mathcal{D}) \propto p(\mathbf{y}|\mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{w}) p(\mathbf{w})$. Then,

$$\begin{aligned}
\log\left(p(\mathbf{y}|\mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{w}) p(\mathbf{w})\right) &= \log\left(\mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1} I) \prod_{i=1}^{N} \mathcal{N}\left(y_i|\mathbf{w}^\top \phi(\mathbf{x}_i), \beta^{-1}\right)\right) \\
&= \log\left(\mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1} I) \mathcal{N}(\mathbf{y}|\Phi \mathbf{w}, \beta^{-1} I)\right) \\
&= -\frac{1}{2}\left(\alpha \mathbf{w}^\top \mathbf{w} + \beta (\mathbf{y} - \Phi \mathbf{w})^T (\mathbf{y} - \Phi \mathbf{w})\right) + (\textit{constant w.r.t. } \mathbf{w}) \\
&= -\frac{1}{2} \mathbf{w}^\top (\alpha I + \beta \Phi^\top \Phi) \mathbf{w} + \mathbf{w}^\top (\beta \Phi^\top \mathbf{y}) + (\textit{constant w.r.t. } \mathbf{w}) \\
&= -\frac{1}{2}\left(\mathbf{w} - \beta (\alpha I + \beta \Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}\right)^\top (\alpha I + \beta \Phi^\top \Phi) \\
&\quad \left(\mathbf{w} - \beta (\alpha I + \beta \Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}\right) + (\textit{constant w.r.t. } \mathbf{w}).
\end{aligned}$$

Thus, the posterior distribution is

$$\mathcal{N}\left(\beta (\alpha I + \beta \Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}, \, (\alpha I + \beta \Phi^\top \Phi)^{-1}\right).$$

Since $\alpha, \beta > 0$ and $\Phi^\top \Phi$ is positive semi-definite, $\alpha I + \beta \Phi^\top \Phi$ is positive definite and $(\alpha I + \beta \Phi^\top \Phi)^{-1}$ is also positive definite. Then, it is clear that the density of a multivariate normal distribution is maximized at the mean, so the MAP estimator of $\mathbf{w}$ is

$$\beta(\alpha I + \beta \Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}.$$

(b) We found the MLE as $\mathbf{w}_{MLE} = (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}$ in the lecture. Therefore,

$$\lim_{\alpha \to 0} \beta(\alpha I + \beta \Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y} = (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}.$$

(c) Let $\mu := \beta(\alpha I + \beta \Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}$ and $\mathbf{\Sigma} := (\alpha I + \beta \Phi^\top \Phi)^{-1}$. Then,

$$\begin{aligned}
p(y_0|\mathcal{D}, \mathbf{x}_0) &= \int p(y_0|\mathbf{w}, \mathbf{x}_0) p(\mathbf{w}|\mathcal{D}) \, d\mathbf{w} \\
&= \int \mathcal{N}\left(y_0|\mathbf{w}^\top \phi(\mathbf{x}_0), \beta^{-1}\right) \mathcal{N}\left(\mathbf{w}|\mu, \mathbf{\Sigma}\right) d\mathbf{w} \\
&\propto \int \exp\left(-\frac{1}{2}\left\{\beta(y_0 - \mathbf{w}^\top \phi(\mathbf{x}_0))^2 + (\mathbf{w} - \mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{w} - \mu)\right\}\right) d\mathbf{w} \\
&\propto \int \exp\left(-\frac{1}{2}\left\{\beta y_0^2 - 2\beta \mathbf{w}^\top \phi(\mathbf{x}_0) y_0 + \beta \mathbf{w}^\top \phi(\mathbf{x}_0)\phi(\mathbf{x}_0)^\top \mathbf{w} + \mathbf{w}^\top \mathbf{\Sigma}^{-1} \mathbf{w} - 2\mathbf{w}^\top \mathbf{\Sigma}^{-1} \mu\right\}\right) d\mathbf{w} \\
&= \exp\left(-\frac{\beta}{2} y_0^2\right) \int \exp\left(-\frac{1}{2} \mathbf{w}^\top \left(\beta \phi(\mathbf{x}_0)\phi(\mathbf{x}_0)^\top + \mathbf{\Sigma}^{-1}\right) \mathbf{w} + \mathbf{w}^\top \left(\beta y_0 \, \phi(\mathbf{x}_0) + \mathbf{\Sigma}^{-1}\mu\right)\right) d\mathbf{w} \\
&\propto \exp\left(-\frac{\beta}{2} y_0^2\right) \exp\left(\frac{1}{2}\left(\beta y_0 \, \phi(\mathbf{x}_0) + \mathbf{\Sigma}^{-1}\mu\right)^\top \left(\beta \phi(\mathbf{x}_0)\phi(\mathbf{x}_0)^\top + \mathbf{\Sigma}^{-1}\right)^{-1} \left(\beta y_0 \, \phi(\mathbf{x}_0) + \mathbf{\Sigma}^{-1}\mu\right)\right) \\
&\times \int \mathcal{N}\left(\mathbf{w} \middle| \left(\beta \phi(\mathbf{x}_0)\phi(\mathbf{x}_0)^\top + \mathbf{\Sigma}^{-1}\right)^{-1} \left(\beta y_0 \, \phi(\mathbf{x}_0) + \mathbf{\Sigma}^{-1}\mu\right), \left(\beta \phi(\mathbf{x}_0)\phi(\mathbf{x}_0)^\top + \mathbf{\Sigma}^{-1}\right)^{-1}\right) d\mathbf{w} \\
&= \exp\left(-\frac{\beta}{2} y_0^2\right) \exp\left(\frac{1}{2}\left(\beta y_0 \, \phi(\mathbf{x}_0) + \mathbf{\Sigma}^{-1}\mu\right)^\top \left(\beta \phi(\mathbf{x}_0)\phi(\mathbf{x}_0)^\top + \mathbf{\Sigma}^{-1}\right)^{-1} \left(\beta y_0 \, \phi(\mathbf{x}_0) + \mathbf{\Sigma}^{-1}\mu\right)\right).
\end{aligned}$$

Let $\mathbf{\Sigma}_0 := \left(\beta \phi(\mathbf{x}_0)\phi(\mathbf{x}_0)^\top + \mathbf{\Sigma}^{-1}\right)^{-1}$. By using the Woodbury matrix identity,

$$\mathbf{\Sigma}_0 = \left(\beta \phi(\mathbf{x}_0)\phi(\mathbf{x}_0)^\top + \mathbf{\Sigma}^{-1}\right)^{-1} = \mathbf{\Sigma} - \frac{\mathbf{\Sigma}\beta \phi(\mathbf{x}_0)\phi(\mathbf{x}_0)^\top \mathbf{\Sigma}}{1 + \beta \phi(\mathbf{x}_0)^\top \mathbf{\Sigma} \, \phi(\mathbf{x}_0)}.$$

Then,

$$\log p(y_0|\mathcal{D}, \mathbf{x}_0) = -\frac{1}{2}\left(\beta - \beta^2 \phi(\mathbf{x}_0)^\top \mathbf{\Sigma}_0 \, \phi(\mathbf{x}_0)\right) y_0^2 + \left(\beta \phi(\mathbf{x}_0)^\top \mathbf{\Sigma}_0 \mathbf{\Sigma}^{-1} \mu\right) y_0 + (\textit{constant w.r.t } y_0).$$

Note that

$$\phi(\mathbf{x}_0)^\top \mathbf{\Sigma}_0 \, \phi(\mathbf{x}_0) = \phi(\mathbf{x}_0)^\top \mathbf{\Sigma} \phi(\mathbf{x}_0) - \frac{\beta \phi(\mathbf{x}_0)^\top \mathbf{\Sigma} \phi(\mathbf{x}_0)\phi(\mathbf{x}_0)^\top \mathbf{\Sigma} \phi(\mathbf{x}_0)}{1 + \beta \phi(\mathbf{x}_0)^\top \mathbf{\Sigma} \, \phi(\mathbf{x}_0)} = \frac{\phi(\mathbf{x}_0)^\top \mathbf{\Sigma} \phi(\mathbf{x}_0)}{1 + \beta \phi(\mathbf{x}_0)^\top \mathbf{\Sigma} \, \phi(\mathbf{x}_0)}$$

and

$$\phi(\mathbf{x}_0)^\top \mathbf{\Sigma}_0 \mathbf{\Sigma}^{-1} = \phi(\mathbf{x}_0)^\top - \frac{\beta \phi(\mathbf{x}_0)^\top \mathbf{\Sigma} \phi(\mathbf{x}_0)\phi(\mathbf{x}_0)^\top}{1 + \beta \phi(\mathbf{x}_0)^\top \mathbf{\Sigma} \, \phi(\mathbf{x}_0)} = \frac{\phi(\mathbf{x}_0)^\top}{1 + \beta \phi(\mathbf{x}_0)^\top \mathbf{\Sigma} \, \phi(\mathbf{x}_0)}.$$

Therefore,

$$\log p(y_0|\mathcal{D}, \mathbf{x}_0) = -\frac{1}{2} \cdot \frac{\beta}{1 + \beta \phi(\mathbf{x}_0)^\top \mathbf{\Sigma} \, \phi(\mathbf{x}_0)} y_0^2 + \frac{\beta \phi(\mathbf{x}_0)^\top \mu}{1 + \beta \phi(\mathbf{x}_0)^\top \mathbf{\Sigma} \, \phi(\mathbf{x}_0)} y_0 + (\textit{constant w.r.t } y_0),$$

which implies $p(y_0|\mathcal{D}, \mathbf{x}_0)$ is a normal distribution with the mean

$$\phi(\mathbf{x}_0)^\top \mu = \beta\phi(\mathbf{x}_0)^\top(\alpha I + \beta\Phi^\top\Phi)^{-1}\Phi^\top \mathbf{y}$$

and the variance

$$\frac{1 + \beta\phi(\mathbf{x}_0)^\top \boldsymbol{\Sigma}\, \phi(\mathbf{x}_0)}{\beta} = \frac{1}{\beta} + \phi(\mathbf{x}_0)^\top(\alpha I + \beta\Phi^\top\Phi)^{-1}\, \phi(\mathbf{x}_0).$$

2. (10pts) Consider a two-class classification problem and the following training set, each having four binary attributes:

class 1	class 2
0110	1011
1010	0000
0011	0100
1111	1110

Construct a (unpruned) decision tree based on the following criteria (ID3):

- Basically, use the information gain as an impurity measure in the splitting criterion. If there are more than one features maximizing the impurity measure, then choose the predecessor. For example, if the first and the second attributes maximize the impurity measure, then choose the first attribute.
- On each iteration, it iterates through every unused attribute of the dataset, i.e. consider only attributes never selected before when the algorithm recurs on each subset.
- The maximum tree depth is 2.

Solution. Note that the orginal entropy is

$$H = -\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8} = 1.$$

Let a_i be the ith attribute and calculate the information gains. Then,

$$\begin{aligned}
G(a_1) &= 1 - \frac{4}{8}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) - \frac{4}{8}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) = 0,\\
G(a_2) &= 1 - \frac{4}{8}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) - \frac{4}{8}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) = 0,\\
G(a_3) &= 1 - \frac{2}{8}\left(-\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2}\right) - \frac{6}{8}\left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right) = \frac{3}{2} - \frac{3}{4}\log_2 3 \approx 0.311,\\
G(a_4) &= 1 - \frac{5}{8}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) - \frac{3}{8}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) = \frac{3}{2} - \frac{5}{8}\log_2 5 \approx 0.049.
\end{aligned}$$

So, the information gain is maximized when we split the data with the third attribute. Let $\mathcal{D}_0$ be the data set whose third attribute is 0 and $\mathcal{D}_1$ be the data set whose third attribute is 1. That is, $\mathcal{D}_0 = \{0000, 0100\}$ and $\mathcal{D}_1 = \{0110, 1010, 0011, 1111, 1011, 1110\}$. Note that since the entropy of $\mathcal{D}_0$ is

$$H = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} = 0,$$

we do not need to split $\mathcal{D}_0$ anymore. Then, we only need to split $\mathcal{D}_1$. The entropy of $\mathcal{D}_1$ is

$$H = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = -\frac{2}{3} + \log_2 3.$$

Then, the information gains for the second iteration are

$$
\begin{aligned}
G(a_1) &= -\frac{2}{3} + \log_2 3 - \frac{2}{6}\left(-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right) - \frac{4}{6}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) = -\frac{4}{3} + \log_2 3 \approx 0.252,\\
G(a_2) &= -\frac{2}{3} + \log_2 3 - \frac{3}{6}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) - \frac{3}{6}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) = 0,\\
G(a_4) &= -\frac{2}{3} + \log_2 3 - \frac{3}{6}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) - \frac{3}{6}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) = 0.
\end{aligned}
$$

So, the information gain is maximized when we split $\mathcal{D}_1$ with the first attribute. Let $\mathcal{D}_{10}$ be the subset of $\mathcal{D}_1$ whose first attribute is 0 and $\mathcal{D}_{11}$ be the subset of $\mathcal{D}_1$ whose first attribute is 1. That is, $\mathcal{D}_{10} = \{0110, 0011\}$ and $\mathcal{D}_{11} = \{1010, 1111, 1011, 1110\}$.

Since the maximum tree depth is 2, we constructed a desired decision tree. Note that $\mathcal{D}_0$ can be classified as class 2, and $\mathcal{D}_{10}$ can be classified as class 1, while $\mathcal{D}_{11}$ is remained undetermined.

3. (5pts) In our lecture, a loss function of a logistic model

$$
y(\mathbf{x}, \mathbf{w}) = \begin{cases} 1 & \text{with probability } \sigma(\mathbf{w}^\top \mathbf{x}) \\ 0 & \text{otherwise} \end{cases}
$$

is given by

$$
E(\mathbf{w}) = -\log \mathcal{L} = -\sum_{i=1}^{N} y_i \log p(y_i = 1 | \mathbf{x}_i, \mathbf{w}) - \sum_{i=1}^{N} (1 - y_i) \log(1 - p(y_i = 1 | \mathbf{x}_i, \mathbf{w}))
$$

where $\mathcal{D} = \{(\mathbf{x}_i, y_i) : i = 1, \cdots, N\}$ is a training set. Prove that $E(\mathbf{w})$ is convex.

Solution.

$$
\begin{aligned}
E(\mathbf{w}) &= -\sum_{i=1}^{N} y_i \log \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}_i}} - \sum_{i=1}^{N} (1 - y_i) \log \frac{e^{-\mathbf{w}^\top \mathbf{x}_i}}{1 + e^{-\mathbf{w}^\top \mathbf{x}_i}}\\
&= \sum_{i=1}^{N} \log(1 + e^{-\mathbf{w}^\top \mathbf{x}_i}) + \sum_{i=1}^{N} (1 - y_i)\mathbf{w}^\top \mathbf{x}_i
\end{aligned}
$$

So,

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{w}} E(\mathbf{w}) &= \sum_{i=1}^{N} \frac{-e^{-\mathbf{w}^\top \mathbf{x}_i} \cdot \mathbf{x}_i}{1 + e^{-\mathbf{w}^\top \mathbf{x}_i}} + \sum_{i=1}^{N} (1 - y_i)\mathbf{x}_i,\\
\frac{\partial^2}{\partial \mathbf{w}^2} E(\mathbf{w}) &= \sum_{i=1}^{N} \frac{e^{-\mathbf{w}^\top \mathbf{x}_i}(1 + e^{-\mathbf{w}^\top \mathbf{x}_i})\mathbf{x}_i \mathbf{x}_i^\top - e^{-2\mathbf{w}^\top \mathbf{x}_i} \cdot \mathbf{x}_i \mathbf{x}_i^\top}{(1 + e^{-\mathbf{w}^\top \mathbf{x}_i})^2} = \sum_{i=1}^{N} \frac{e^{-\mathbf{w}^\top \mathbf{x}_i} \cdot \mathbf{x}_i \mathbf{x}_i^\top}{(1 + e^{-\mathbf{w}^\top \mathbf{x}_i})^2}
\end{aligned}
$$

Since $\mathbf{x}_i \mathbf{x}_i^\top$ is positive semi-definite and $\frac{e^{-\mathbf{w}^\top \mathbf{x}_i}}{(1 + e^{-\mathbf{w}^\top \mathbf{x}_i})^2} > 0$, the Hessian matrix is also positive semi-definite. Thus, $E(\mathbf{w})$ is convex.

4. (10pts) Consider a dataset

$$
\mathcal{D} = \left\{([-2, -2]^\top, -1), ([-1, 1]^\top, -1), ([3, -1]^\top, -1), ([2, 5]^\top, 1), ([3, 4]^\top, 1), ([3, 5]^\top, 1)\right\}.
$$

Compute the closed form of a linear (hard-margin) SVM classifier. Which points are the support vectors? Verify your answer.

Solution. Let $y = ax + b$ be the linear boundary. Note that the line of the form $x = k$ cannot be the decision boundary. We divide the cases with respect to a.

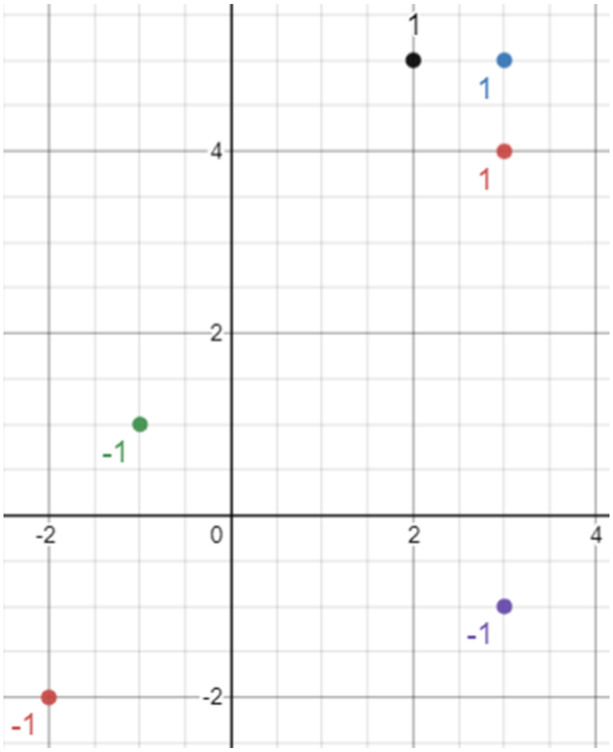

Figure 1: points in a dataset

(a) If $a \geq \frac{3}{4}$ or $a \leq -6$, then we cannot classify the dataset.

(b) If $-\frac{1}{2} \leq a < \frac{3}{4}$, then $(-1, 1)$ and $(3, 4)$ both become support vectors (if $a = -\frac{1}{2}$, $(3, -1)$ is also a support vector) and the margin becomes a distance between $y = ax + 1 + a$ and $(3, 4)$. With simple calculation, we can see that the margin is maximized when $a = -\frac{1}{2}$. In this case the margin is $2\sqrt{5}$.

(c) If $-1 \leq a < -\frac{1}{2}$, then $(3, -1)$ and $(3, 4)$ both become support vectors (if $a = -1$, $(2, 5)$ is also a support vector) and the margin becomes a distance between $y = ax - 1 - 3a$ and $(3, 4)$. With simple calculation, we can see that the margin is increasing as a increases. So, the margin is smaller than $2\sqrt{5}$ in this case.

(d) If $-6 < a < -1$, then $(3, -1)$ and $(2, 5)$ both become support vectors. Similarly, the margin is a distance between $y = ax + 3 + a$ and $(2, 5)$ and is increasing as a increases. Since the distance between $y = ax + 3 + a$ and $(2, 5)$ is $\frac{5}{\sqrt{2}}$ when $a = -1$, the margin is smaller than $2\sqrt{5}$ in this case.

Therefore, the decision boundary should be of the form $y = -\frac{1}{2}x + b$. Since $y = -\frac{1}{2}x + \frac{1}{2}$ and $y = -\frac{1}{2}x + \frac{11}{2}$ contain support vectors and are parallel to the decision boundary, we get $b = \frac{1}{2}(\frac{1}{2} + \frac{11}{2}) = 3$.

Thus, $\mathrm{sign}(x + 2y - 6)$ is the closed form of a linear SVM classifier where $[x, y]^\top$ is an input. Note that

$$\mathrm{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases} .$$

Also, $\{[-1, 1]^\top, [3, -1]^\top, [3, 4]^\top\}$ are support vectors.

5. (10pts) Consider a synthetic training set `train_data` and validation set `valid_data` which are generated from the following code.

```
import numpy as np
import pandas as pd
from sklearn.datasets import make_circles

train_input, train_label = make_circles(n_samples = 150, factor = 0.6, noise = 0.1, random_state = 5)
train_data = pd.DataFrame(train_input, columns=['X', 'Y'])
train_data['label'] = np.array(train_label)

valid_input, valid_label = make_circles(n_samples = 50, factor = 0.6, noise = 0.1, random_state = 8)
valid_data = pd.DataFrame(valid_input, columns=['X', 'Y'])
valid_data['label'] = np.array(valid_label)
```

Inputs of dataset are 2-dimensional vectors (`X` and `Y`) and labels of dataset (`label`) are 0 or 1. In this problem, we construct a binary classifier for this dataset.

(a) Learn (soft-margin) linear SVM models with polynomial features of degree 3 and hyperparameter $C = 1, 5, 10, 50$ (in lecture note) using the given training set.

(b) Learn kernelized SVM models with the RBF kernel

$$K(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x} - \mathbf{z}\|^2)$$

and hyperparameter $C = 1, 5, 10, 50$ (in lecture note) using the given training set.

(c) Calculate the accuracy of our models using the given validation set. Which model is the best?

Solution. Python code and the results are in the .ipynb file. In this experiment, the soft-margin linear SVM model with polynomial features of degree 3 and hyperparameter $C = 1$ and the kernelized SVM model with the RBF kernel $K(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x} - \mathbf{z}\|^2)$ and hyperparameter $C = 10$ are the best models. (accuracy $= 0.98$) It can be different due to the randomness.