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 $\mathcal{D} = \{(\mathbf{x}_i, y_i) : 1 \le i \le N\}$ 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

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

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.

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.

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.

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?