COGS 118A Introduction to Machine Learning I
Midterm Exam II Practice Problems

Instructor: Prof. Zhuowen Tu

1. Shattering

Use shattering to derive the VC-dimension for classifiers below. Show your work.

1) $f(x, w, b) = \text{sign}\big(wx + b\big)$

2) $f(x, q, b) = \text{sign}\big(qx^2 + b\big)$

3) $f(x, w, b) = \text{sign}\big((wx + b)^2\big)$

where $x, w, q, b \in \mathbb{R}$ and w, q, b are free parameters. Besides, sign(x) is defined as

$$\text{sign}(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0. \end{cases}$$

Ans:
See HW5 Solution.

2. Logistic Regression

In logistic regression model, y is the label for each data point, and it can be either 0 or 1. Here, we define our approximate function to

$$p(y=1|\mathbf{x}) = h(\mathbf{x};\mathbf{w},b) = \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+b)}}$$

where $\mathbf{x} \in \mathbb{R}^m$ is the feature, and b is the bias, and $\mathbf{w} \in \mathbb{R}^K$ contains the set of parameters $\mathbf{w} = (w_0, w_1, ..., w_m)$. The output of $h(\mathbf{x};\mathbf{w},b)$ is called the confidence. When $h(\mathbf{x};\mathbf{w},b) >= 0.5$, the classifier outputs 1 for the given $\mathbf{x}$; otherwise, the classifier outputs 0.

The goal is to train a classifier based on logistic regression to predict the correct label y_i from the given feature $\mathbf{x}_i$. We define a loss function which measures the distance between the correct label and the prediction from the classifier, as is shown below:

$$\mathcal{L}(\mathbf{w}) = -\sum_i \ln p(y_i|\mathbf{x}_i;\mathbf{w},b)$$

where

$$p(y_i|\mathbf{x}_i;\mathbf{w}+b) = \frac{1}{1+e^{-(2y_i-1)\times(\mathbf{w}^T\mathbf{x}_i+b)}},$$

which is called the sigmoid function.

The training procedure is to minimize the loss function on the training set. Consider ***gradient descent*** method to find the optimal $\mathbf{w}^*$ in $h(\mathbf{x};\mathbf{w},b)$.

(a) Derive $\dfrac{\partial\mathcal{L}(\mathbf{w})}{\partial\mathbf{w}}$.

(b) Suppose that the learning rate is denoted as α. Write down the update rule for $\mathbf{w}$.

Ans:
See HW4 Solution.

3. Linear Discriminative Analysis

Consider a dataset which has two classes $y \in \{-1, +1\}$. For $y = -1$, there are two data points $\mathbf{x}_1 = [2, 5]^\top, \mathbf{x}_2 = [2, 1]^\top$; for $y = +1$, there are two data points $\mathbf{x}_3 = [5, 3]^\top, \mathbf{x}_4 = [7, 3]^\top$. Now we have two options to project the data points, $\mathbf{w}_1 = [1, 0]^\top$ and $\mathbf{w}_2 = [0, 1]^\top$. Please compare which projection is better based on the Fisher's linear discriminant analysis.

Hint: The Fisher's linear discriminant analysis is defined to maximize the criterion function:

$$S(\mathbf{w}) = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{\left(\mathbf{w}^\top \mu_0 - \mathbf{w}^\top \mu_1\right)^2}{\mathbf{w}^\top (\Sigma_1 + \Sigma_0) \mathbf{w}}$$

Ans:

We can get $\mu_0 = [2, 3]^\top$, $\mu_1 = [6, 3]^\top$.
Then for $\mathbf{w}_2 = [0, 1]^\top$, $S(\mathbf{w}) = 0$; for $\mathbf{w}_1 = [1, 0]^\top$, $S(\mathbf{w}) > 0$.
So $\mathbf{w}_1 = [1, 0]^\top$ is better projection according to Fisher's linear discriminant analysis.
It is also recommended to plot these points to see why $\mathbf{w}_1 = [1, 0]^\top$ is a better projection.

4. Support Vector Machine

Consider a dataset $\{(\mathbf{x}^{(i)}, y^{(i)})\}$ where $\mathbf{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}]^\top$ and $y^{(i)} \in \{-1, +1\}$, which is shown in the figure below. Suppose we have trained a support vector machine (SVM) on the dataset, which has the **decision boundary** $\mathbf{w} \cdot \mathbf{x} + b = 0$, **positive plane** $\mathbf{w} \cdot \mathbf{x} + b = +1$ and **negative plane** $\mathbf{w} \cdot \mathbf{x} + b = -1$. The SVM is optimized as following:

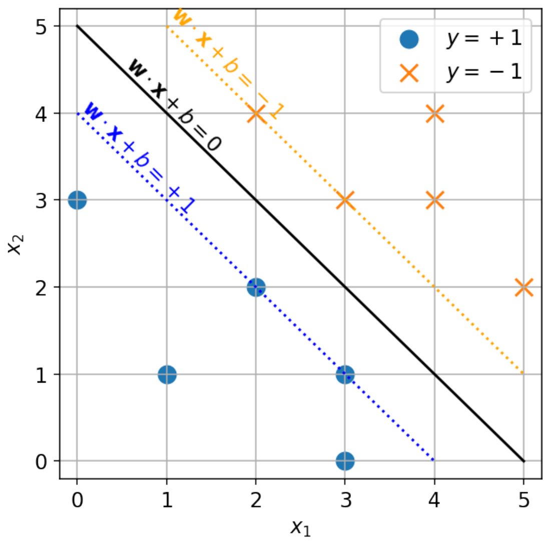
$$\begin{aligned} \text{Find: } & \arg\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 \\ \text{Subject to: } & \mathbf{w} \cdot \mathbf{x}^{(i)} + b \geq +1, \text{ if } y^{(i)} = +1 \\ & \mathbf{w} \cdot \mathbf{x}^{(i)} + b \leq -1, \text{ if } y^{(i)} = -1. \end{aligned}$$

1) Please draw the **decision boundary**, **positive plane** and **negative plane** in the figure below.

2) Calculate the $\mathbf{w}$ and b from your drawn **positive plane** and **negative plane**.

3) Calculate the size of the margin.

4) Calculate the hinge loss for data points: $([3,4]^\top, 1), ([2,1]^\top, 1), ([2,1]^\top, -1), ([2,3]^\top, -1)$.



Ans:

1) The decision boundary is plotted as above. You should be able to draw the decision boundary that maximizes the margin.

2) $\mathbf{w} = [-1, -1]^\top, b = 5$.

3) The size of the margin can be computed using $\frac{2}{||\mathbf{w}||} = \sqrt{2}$.

4) The hinge loss of $([3,4]^\top, 1), ([2,1]^\top, 1), ([2,1]^\top, -1), ([2,3]^\top, -1)$ are 3, 0, 3, 1 respectively.

5. Regression

You are given data $S = \{(x_i, y_i), i = 1, \ldots, n\}$. The data are expressed as matrices $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^\top$ and $Y = [y_1, y_2, \ldots, y_n]^\top$ where $\mathbf{x}_i$ is $[1, x_i, x_i^2]^\top$. The parabola function is defined as: $f(x; W) = w_0 + w_1 x + w_2 x^2 = \mathbf{x}^\top W$ where $\mathbf{x}$ is $[1, x, x^2]^\top$ and W is $[w_0, w_1, w_2]^\top$. We consider a loss function with combined L_1 loss term and L_2 loss term as follows:

$$\mathcal{L} = \sum_{i=1}^{N} \alpha \frac{1}{2}(y_i - f(x_i; W))^2 + (1 - \alpha)|y_i - f(x_i; W)|$$

where α is the weight of L_2 loss term. We wish to optimize W using gradient descent.

(a) Obtain the gradient $\frac{\partial g(W)}{\partial W}$.

(b) If λ denotes learning rate, what is the update rule for W?

Ans:

See HW4 Solution.

6. Cross Validation

Given a dataset $S = \{(x^{(i)}, y^{(i)})\}$ where $x^{(i)} \in \mathbb{R}$, $y^{(i)} \in \{-1, +1\}$ and $|S| = 6$. The items in the dataset are given below:

$$(x^{(1)}, y^{(1)}) = (1, -1), \quad (x^{(2)}, y^{(2)}) = (2, -1), \quad (x^{(3)}, y^{(3)}) = (3, +1),$$
$$(x^{(4)}, y^{(4)}) = (4, -1), \quad (x^{(5)}, y^{(5)}) = (5, +1), \quad (x^{(6)}, y^{(6)}) = (6, +1).$$

Suppose you are training a linear classifier $f(x; a, b) = \text{sign}(ax+b)$ with 2-fold cross validation:

1) Split the dataset S into:

$$S_1 = \{(x^{(1)}, y^{(1)}), (x^{(3)}, y^{(3)}), (x^{(4)}, y^{(4)})\}$$
$$S_2 = \{(x^{(2)}, y^{(2)}), (x^{(5)}, y^{(5)}), (x^{(6)}, y^{(6)})\}$$

2) Train the classifier $f(x; a, b)$ on S_1, get the parameters $a_1 = 2, b_1 = -5$ and then validate the classifier on S_2.

3) Train the classifier $f(x; a, b)$ on S_2, get the parameters $a_2 = 2, b_2 = -7$ and then validate the classifier on S_1.

Please finish the tasks below:

(a) Calculate the **average training accuracy** in the 2-fold cross validation.

(b) Calculate the **average validation accuracy** in the 2-fold cross validation.

Ans:
a) Predicts the labels of the training set $S1$:
$\hat{y_1} = -1, \hat{y_3} = 1, \hat{y_4} = 1$, so the training accuracy is 2/3.
Predicts the labels of the training set $S2$:
s $\hat{y_2} = -1, \hat{y_5} = 1, \hat{y_6} = 1$, so the training accuracy is 1.
The average training accuracy is 5/6
b) Predicts the labels of the validation set $S2$:
$\hat{y_2} = -1, \hat{y_5} = 1, \hat{y_6} = 1$, so the validation accuracy is 1.
Predicts the labels of the validation set $S1$:
$\hat{y_1} = -1, \hat{y_3} = -1, \hat{y_4} = 1$, so the validation accuracy is 1/3.
The average validation accuracy is 2/3