Intro to Big Data Science: Assignment 3

Due Date: Apr 8, 2025

✏ **Exercise 1 (Decision Tree)**

You are trying to determine whether a boy finds a particular type of food appealing based on the food's temperature, taste, and size.

Food Sample Id	Appealing	Temperature	Taste	Size
1	No	Hot	Salty	Small
2	No	Cold	Sweet	Large
3	No	Cold	Sweet	Large
4	Yes	Cold	Sour	Small
5	Yes	Hot	Sour	Small
6	No	Hot	Salty	Large
7	Yes	Hot	Sour	Large
8	Yes	Cold	Sweet	Small
9	Yes	Cold	Sweet	Small
10	No	Hot	Salty	Large

1. What is the initial entropy of "Appealing"?
2. Assume that "Taste" is chosen as the root of the decision tree. What is the information gain associated with this attribute.
3. Draw the full decision tree learned from this data (without any pruning).

✏ **Exercise 2 (k-Nearest-Neighbors)** Suppose you have 10,000 data points $\{(x_k, y_k) : k = 1, 2, \ldots, 10000\}$. Your dataset has one input and one output. The k-th data point is generated by the following recipe:

$$x_k = k/10000, y_k \sim N(0, 2^2).$$

So that y_k is all noise: drawn from a Gaussian with mean 0 and variance $\sigma^2 = 4$. Note that its value is independent of all other y values. You are considering two learning algorithms:

- Algorithm NN: 1-nearest neighbor.
- Algorithm Zero: Always predict zero.

1. What is the **expected mean squared training error** for Algorithm NN?
2. What is the **expected mean squared training error** for Algorithm Zero?
3. Recall the leave-one-out cross validation estimator is defined over the training set $\{(x_k, y_k) : k = 1, 2, \ldots, 10000, k \neq i\}$ for each sample (x_i, y_i). What is the **expected mean squared leave-one-out cross-validation error** for Algorithm NN?
4. What is the **expected mean squared leave-one-out cross-validation error** for Algorithm Zero?

✏ **Exercise 3 (Naive Bayes)** Suppose you have the following training set with three boolean inputs x, y and z, and a boolean output U. Suppose you have to predict U using a naive Bayes classifier. Then after learning is complete what would be the predicted probability $P(U = 0|x = 1, y = 0, z = 0)$?

x	y	z	U
1	1	1	0
1	1	0	0
0	0	0	0
0	1	0	1
1	0	1	1
0	1	1	1

✏ **Exercise 4 (Soft-Margin Linear Support Vector Machine)**

Given the following dataset aligning on the x-axis (See the figure below), which consists of 4 positive data points $\{0, 1, 2, 3\}$ and 3 negative data points $\{-3, -2, -1\}$. Suppose that we want to learn a soft-margin linear SVM for this data set. Remember that the soft-margin linear SVM can be formalized as the following constrained quadratic optimization problem. In this formulation, C is the regularization parameter, which balances the size of margin (i.e., smaller $\|\mathbf{w}\|_2^2$) vs. the violation of the margin (i.e., smaller $\sum_{i=1}^m \xi_i$).

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^n \xi_i$$
$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geqslant 1 - \xi_i, \quad \xi_i \geqslant 0, \quad i = 1, \ldots, n$$

1. If $C = 0$, which means that we only care the size of the margin, how many support vectors do we have?
2. if $C \to \infty$, which means that we only care the violation of the margin, how many support vectors do we have?

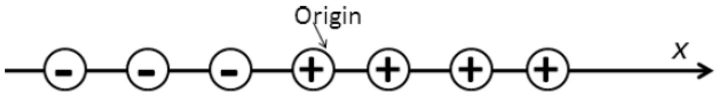

Figure 1: The data set.

3. Properties of Kernel:
   a) Using the definition of kernel functions in SVM, prove that the kernel $K(\mathbf{x}_i, \mathbf{x}_j)$ is symmetric, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the feature vectors for i-th and j-th examples.
   b) Given n training examples $(\mathbf{x}_i, \mathbf{x}_j)$ for $(i, j = 1, \ldots, n)$, the kernel matrix A is an $n \times n$ square matrix, where $A(i, j) = K(\mathbf{x}_i, \mathbf{x}_j)$. Prove that the kernel matrix A is semi-positive definite.

✏ **Exercise 5 (Error bound for 1-nearest-neighbor method, optional)** In class, we have estimated that the error for 1-nearest-neighbor rule is roughly twice the Bayes error. Now let us make it more rigorous.

Let us consider the two-class classification problem with $\mathscr{X} = [0,1]^d$ and $\mathscr{Y} = \{0,1\}$. The underlying joint probability distribution on $\mathscr{X} \times \mathscr{Y}$ is $P(\mathbf{X}, Y)$ from which we deduce that the marginal distribution of $\mathbf{X}$ is $p_{\mathbf{X}}(\mathbf{x})$ and the conditional probability distribution is $\eta(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$. Assume that $\eta(\mathbf{x})$ is c-Lipschitz continuous: $|\eta(\mathbf{x}) - \eta(\mathbf{x}')| \leqslant c\|\mathbf{x} - \mathbf{x}'\|$ for any $\mathbf{x}, \mathbf{x}' \in \mathscr{X}$. Recall that the Bayes rule is $f^*(\mathbf{x}) = 1_{\{\eta(\mathbf{x})>1/2\}}$. Given a training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $(\mathbf{x}_i, y_i) \overset{i.i.d.}{\sim} P$ (or equivalently $S \sim P^n$), the 1-nearest-neighbor rule is $f^{1NN}(\mathbf{x}) = y_{\pi_S(\mathbf{x})}$ where $\pi_S(\mathbf{x}) = \arg\min_i \|\mathbf{x} - \mathbf{x}_i\|$.

Define the generalization error for rule f as $\mathscr{E}(f) = \mathrm{E}_{(\mathbf{X},Y)\sim P} 1_{Y \neq f(\mathbf{X})}$. Show that

$$\mathrm{E}_{S\sim P^n}\mathscr{E}(f^{1NN}) \leqslant 2\mathscr{E}(f^*) + c\mathrm{E}_{S\sim P^n}\mathrm{E}_{\mathbf{x}\sim p_{\mathbf{X}}}\|\mathbf{x} - \mathbf{x}_{\pi_S(\mathbf{x})}\|.$$

(This means that we can have a precise error estimate for 1-nearest-neighbor rule if we can bound $\mathrm{E}_{S\sim P^n}\mathrm{E}_{\mathbf{x}\sim p_{\mathbf{X}}}\|\mathbf{x} - \mathbf{x}_{\pi(\mathbf{x})}\|$.)

✏ **Exercise 6** Online study and exercises.