Assignment 3 Solutions

Hesipeng

June 8, 2025

Exercise 1 (Decision Tree)

1. Initial Entropy of "Appealing"

Given the dataset with 10 samples (5 "Yes" and 5 "No"), the initial entropy is calculated as:

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i = -\left(\frac{5}{10}\log_2\frac{5}{10} + \frac{5}{10}\log_2\frac{5}{10}\right) = 1$$

2. Information Gain for "Taste"

The information gain when splitting on "Taste" is:

Table 1: Entropy calculation for each taste value

Taste	Count	Yes/No	Entropy
Salty	3	0/3	0
Sweet	4	3/1	1
Sour	3	3/0	0

$$H(S|Taste) = \frac{3}{10}\times 0 + \frac{4}{10}\times 1 + \frac{3}{10}\times 0 = 0.4$$
$$IG(Taste) = H(S) - H(S|Taste) = 1 - 0.4 = 0.6$$

3. Decision Tree Structure

The decision tree can be represented as:

```
Root (Taste)
 Salty → No
 Sour → Yes
 Sweet
     Small → Yes
     Large → No
```

Exercise 2 (k-Nearest-Neighbors)

1. Expected MSE for 1-NN

For 1-NN, the training error is always 0:

$$\text{MSE}_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$$

2. Expected MSE for Zero Predictor

For the zero predictor:

$$\text{MSE}_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - 0)^2 = \text{Var}(y) = 4$$

3. Expected LOO-CV MSE for Algorithm NN

For each left-out point, the prediction comes from its nearest neighbor (which is independent since y's are independent):

$$\text{MSE}_{\text{LOO}} = \frac{1}{10000}\sum_{k=1}^{10000} \text{E}[(y_k - y_j)^2] = \text{E}[y_k^2] + \text{E}[y_j^2] = 4 + 4 = 8$$

4. Expected LOO-CV MSE for Algorithm Zero

Same as training MSE since predictions don't depend on the data:

$$\text{MSE}_{\text{LOO}} = 4$$

Exercise 3 (Naive Bayes)

Given the query point $(x = 1, y = 0, z = 0)$, we calculate: Total samples: 6

$$P(U = 0) = \frac{3}{6} = 0.5 \quad P(U = 1) = \frac{3}{6} = 0.5$$

For $U = 0$ (3 samples):

$$P(x = 1|U = 0) = \frac{2}{3} \quad P(y = 0|U = 0) = \frac{1}{3} \quad P(z = 0|U = 0) = \frac{2}{3}$$

For $U = 1$ (3 samples):

$$P(x = 1|U = 1) = \frac{1}{3} \quad P(y = 0|U = 1) = \frac{1}{3} \quad P(z = 0|U = 1) = \frac{1}{3}$$

$$\begin{aligned} P(U = 0|x = 1, y = 0, z = 0) &\propto P(U = 0) \cdot P(x = 1|U = 0) \cdot P(y = 0|U = 0) \cdot P(z = 0|U = 0) \\ &= 0.5 \times \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} \\ &= \frac{0.5 \times 4}{27} = \frac{2}{27} \end{aligned}$$

$$\begin{aligned} P(U = 1|x = 1, y = 0, z = 0) &\propto P(U = 1) \cdot P(x = 1|U = 1) \cdot P(y = 0|U = 1) \cdot P(z = 0|U = 1) \\ &= 0.5 \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \\ &= \frac{0.5 \times 1}{27} = \frac{1}{54} \end{aligned}$$

$$\text{Normalization factor} = \frac{2}{27} + \frac{1}{54} = \frac{5}{54}$$

$$P(U = 0|x = 1, y = 0, z = 0) = \frac{\frac{2}{27}}{\frac{5}{54}} = \frac{4}{5}$$

$$P(U = 1|x = 1, y = 0, z = 0) = \frac{\frac{1}{54}}{\frac{5}{54}} = \frac{1}{5}$$

The predicted probability is:

$$P(U = 0|x = 1, y = 0, z = 0) = \boxed{\frac{4}{5}}$$

Exercise 4 (SVM)

1. Support Vectors when C=0

When $C = 0$, all points become support vectors:

$$\text{Number of SVs} = 7 \quad \text{(all data points)}$$

2. Support Vectors when C$\to\infty$

When $C \to \infty$, only the boundary points are support vectors:

$$\text{Number of SVs} = 2 \quad \text{(points at the decision boundary)}$$

3. Kernel Properties

Proof of Symmetry. For any kernel K and vectors $\mathbf{x}_i, \mathbf{x}_j$:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}_i) \rangle = K(\mathbf{x}_j, \mathbf{x}_i)$$

$\square$

Proof of PSD. For any vector $\mathbf{v}$:

$$\begin{aligned}\mathbf{v}^T A \mathbf{v} &= \sum_{i,j} v_i K(\mathbf{x}_i, \mathbf{x}_j) v_j \\ &= \left\| \sum_i v_i \phi(\mathbf{x}_i) \right\|^2 \geq 0\end{aligned}$$

$\square$

Exercise 5 (Error bound for 1-nearest-neighbor method)

We aim to prove the error bound for 1-nearest-neighbor classification:

$$\mathrm{E}_{S\sim P^n} \mathcal{E}(f^{1NN}) \leq 2\mathcal{E}(f^*) + c\mathrm{E}_{S\sim P^n} \mathrm{E}_{\mathbf{x}\sim p_X} \|\mathbf{x} - \mathbf{x}_{\pi_S(\mathbf{x})}\|$$

Proof

Proof. We decompose the error of the 1-NN classifier as follows:

$$\begin{aligned}
\mathcal{E}(f^{1NN}) &= \mathrm{E}_{(\mathbf{X},Y)\sim P} 1_{Y \neq f^{1NN}(\mathbf{X})} \\
&= \mathrm{E}_{\mathbf{X}\sim p_X} \mathrm{E}_{Y|\mathbf{X}} [1_{Y \neq y_{\pi_S(\mathbf{x})}}] \\
&= \mathrm{E}_{\mathbf{X}\sim p_X} [\eta(\mathbf{x})(1 - y_{\pi_S(\mathbf{x})}) + (1 - \eta(\mathbf{x})) y_{\pi_S(\mathbf{x})}]
\end{aligned}$$

The Bayes error is:

$$\mathcal{E}(f^*) = \mathrm{E}_{\mathbf{X}\sim p_X} [\min(\eta(\mathbf{x}), 1 - \eta(\mathbf{x}))]$$

Now consider the excess error:

$$\begin{aligned}
&\mathrm{E}_{(\mathbf{X},Y)\sim P} [1_{Y \neq y_{\pi_S(\mathbf{x})}} - 1_{Y \neq f^*(\mathbf{x})}] \\
&= \mathrm{E}_{\mathbf{X}\sim p_X} [\eta(\mathbf{x})(1 - 2y_{\pi_S(\mathbf{x})}) + (1 - \eta(\mathbf{x}))(2y_{\pi_S(\mathbf{x})} - 1)] 1_{f^*(\mathbf{x}) \neq y_{\pi_S(\mathbf{x})}} \\
&= \mathrm{E}_{\mathbf{X}\sim p_X} [|2\eta(\mathbf{x}) - 1| 1_{f^*(\mathbf{x}) \neq y_{\pi_S(\mathbf{x})}}]
\end{aligned}$$

Using the Lipschitz condition $|\eta(\mathbf{x}) - \eta(\mathbf{x}_{\pi_S(\mathbf{x})})| \leq c\|\mathbf{x} - \mathbf{x}_{\pi_S(\mathbf{x})}\|$, we have: When $f^*(\mathbf{x}) \neq y_{\pi_S(\mathbf{x})}$:

$$\begin{aligned}
|2\eta(\mathbf{x}) - 1| &\leq |2\eta(\mathbf{x}) - 2\eta(\mathbf{x}_{\pi_S(\mathbf{x})})| + |2\eta(\mathbf{x}_{\pi_S(\mathbf{x})}) - 1| \\
&\leq 2c\|\mathbf{x} - \mathbf{x}_{\pi_S(\mathbf{x})}\| + 1_{f^*(\mathbf{x}) \neq f^*(\mathbf{x}_{\pi_S(\mathbf{x})})}
\end{aligned}$$

The second term contributes exactly $\mathcal{E}(f^*)$ when $f^*(\mathbf{x}) \neq f^*(\mathbf{x}_{\pi_S(\mathbf{x})})$. Therefore:

$$\begin{aligned}
\mathrm{E}_{S\sim P^n} \mathcal{E}(f^{1NN}) &\leq \mathcal{E}(f^*) + \mathrm{E}_{S\sim P^n} \mathrm{E}_{\mathbf{x}\sim p_X} [|2\eta(\mathbf{x}) - 1| 1_{f^*(\mathbf{x}) \neq y_{\pi_S(\mathbf{x})}}] \\
&\leq \mathcal{E}(f^*) + \mathrm{E}_{S\sim P^n} \mathrm{E}_{\mathbf{x}\sim p_X} [2c\|\mathbf{x} - \mathbf{x}_{\pi_S(\mathbf{x})}\| + 1_{f^*(\mathbf{x}) \neq f^*(\mathbf{x}_{\pi_S(\mathbf{x})})}] \\
&\leq 2\mathcal{E}(f^*) + 2c \mathrm{E}_{S\sim P^n} \mathrm{E}_{\mathbf{x}\sim p_X} \|\mathbf{x} - \mathbf{x}_{\pi_S(\mathbf{x})}\|
\end{aligned}$$

This completes the proof. □

Interpretation

The bound shows that the 1-NN error is bounded by:

- Twice the Bayes error (the irreducible error)
- A term proportional to the expected nearest neighbor distance

The constant c represents the smoothness of the conditional probability function $\eta(\mathbf{x})$. Smoother problems (smaller c) will have better 1-NN performance.