

# Homework 1

## CSC2515 Winter 2026

David

January 28, 2026

### Question 1: Training and Testing Error Curves

This question illustrates the phenomena of underfitting and overfitting through idealized learning curves.

#### (a) Error vs. Model Complexity

Figure 1 shows training and testing error as a function of model complexity. As complexity increases, the training error decreases monotonically, since more expressive models can fit the training data more closely. In contrast, the testing error initially decreases as bias is reduced, reaches a minimum at an optimal complexity, and then increases due to overfitting and high variance.

The Bayes error represents the irreducible error inherent in the data-generating process and serves as a lower bound on achievable performance.

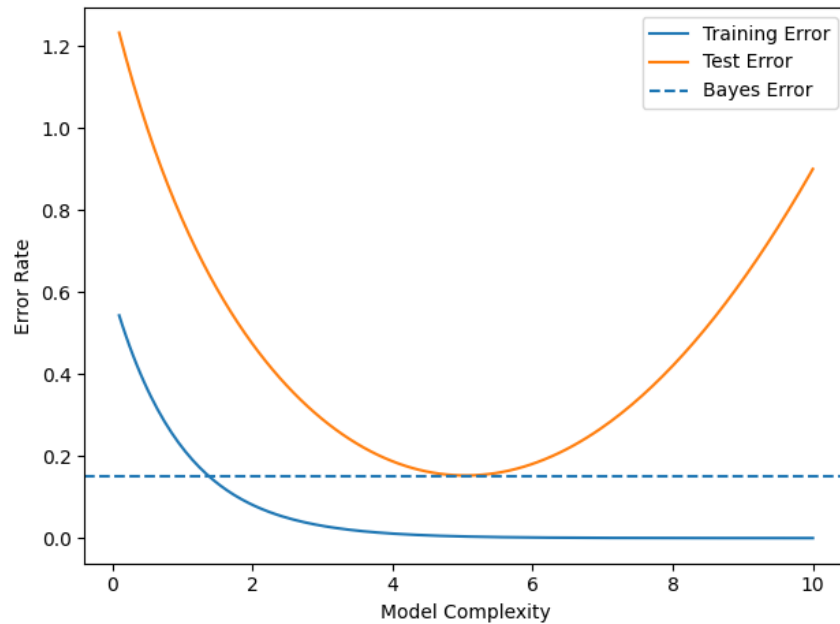


Figure 1: Training and testing error as a function of model complexity. The U-shaped test error curve illustrates the bias-variance tradeoff.

## (b) Error vs. Training Set Size

Figure 2 depicts training and testing error as functions of the training set size for a fixed model complexity. As more data becomes available, the training error typically increases slightly, since fitting all data points perfectly becomes more difficult. Meanwhile, the testing error decreases as the model generalizes better.

Both curves converge to the same asymptotic error level, which lies above the Bayes error. This reflects the fact that, even with infinite data, a fixed-capacity model cannot outperform the Bayes-optimal classifier.

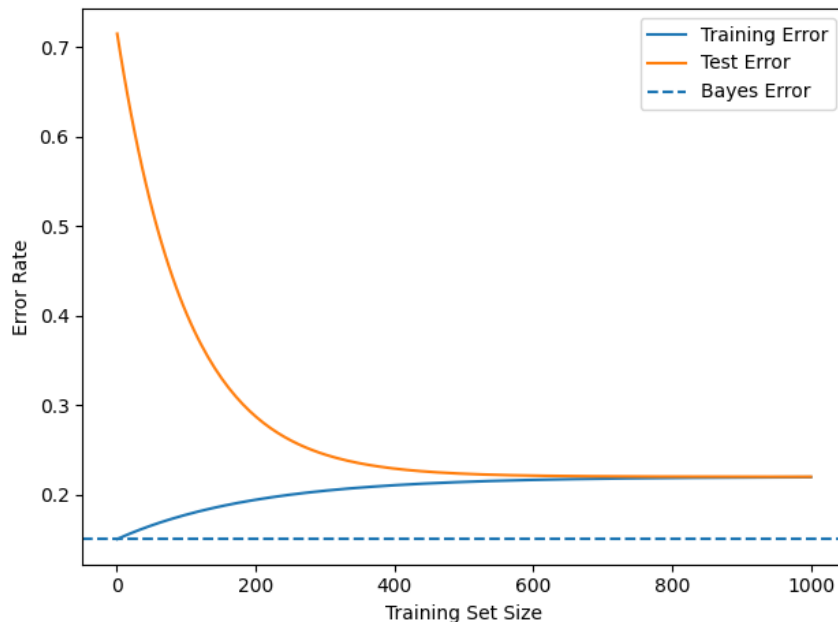


Figure 2: Training and testing error as a function of training set size. Increasing data reduces variance and improves generalization.

## Question 2: Nearest Neighbours & High Dimensions

This question analyzes the behavior of squared Euclidean distances in high-dimensional spaces, illustrating the curse of dimensionality.

### (a) One-dimensional case

Let  $X, Y \sim \text{Unif}[0, 1]$  be independent random variables, and define

$$Z = (X - Y)^2.$$

We first compute the expectation. By symmetry,

$$\mathbb{E}[Z] = \int_0^1 \int_0^1 (x - y)^2 dx dy.$$

Evaluating the inner integral,

$$\int_0^1 (x - y)^2 dx = \int_0^1 (x^2 - 2xy + y^2) dx = \frac{1}{3} - y + y^2.$$

Integrating over  $y$  gives

$$\mathbb{E}[Z] = \int_0^1 \left( \frac{1}{3} - y + y^2 \right) dy = \frac{1}{6}.$$

Next, we compute the second moment:

$$\mathbb{E}[Z^2] = \mathbb{E}[(X - Y)^4] = \int_0^1 \int_0^1 (x - y)^4 dx dy.$$

Evaluating yields

$$\mathbb{E}[Z^2] = \frac{1}{15}.$$

Therefore,

$$\text{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 = \frac{1}{15} - \left( \frac{1}{6} \right)^2 = \frac{7}{180}.$$

### (b) $d$ -dimensional case

Let  $X, Y \in [0, 1]^d$  with independent coordinates, and define the squared Euclidean distance

$$R = \|X - Y\|_2^2 = \sum_{i=1}^d (X_i - Y_i)^2.$$

Each term  $(X_i - Y_i)^2$  is an independent copy of  $Z$  from part (a). By linearity of expectation,

$$\mathbb{E}[R] = \sum_{i=1}^d \mathbb{E}[Z] = \frac{d}{6}.$$

Similarly, using independence,

$$\text{Var}(R) = \sum_{i=1}^d \text{Var}(Z) = \frac{7d}{180}.$$

### (c) Interpretation in high dimensions

The maximum possible squared distance between two points in the unit cube  $[0, 1]^d$  is attained at opposite corners and equals  $d$ . In contrast, the expected squared distance satisfies

$$\mathbb{E}[R] = \frac{d}{6}.$$

Thus, typical pairwise distances are a constant fraction of the maximum distance. Moreover, since the standard deviation of  $R$  scales as  $\sqrt{d}$  while the mean scales as  $d$ , the relative variability  $\sqrt{\text{Var}(R)}/\mathbb{E}[R]$  decays like  $1/\sqrt{d}$ . This concentration phenomenon implies that, in high dimensions, most points are approximately the same distance apart, illustrating the curse of dimensionality for nearest-neighbor methods.

## Question 3: Information Theory Properties

Throughout, let  $X, Y$  be discrete random variables with probability mass functions  $p(x)$  and  $q(x)$ .

### (a) Entropy Non-negativity

**Theorem.** For any discrete random variable  $X$ , the entropy satisfies  $H(X) \geq 0$ .

*Proof.* By definition,

$$H(X) = - \sum_x p(x) \log p(x).$$

For all  $x$  with  $p(x) > 0$ , we have  $0 < p(x) \leq 1$ , hence  $\log p(x) \leq 0$  and so  $-\log p(x) \geq 0$ . Each term in the sum is therefore non-negative, implying  $H(X) \geq 0$ .  $\square$

### (b) Chain Rule for Entropy

**Theorem.** The joint entropy satisfies

$$H(X, Y) = H(X) + H(Y | X).$$

*Proof.* By definition of joint entropy,

$$H(X, Y) = - \sum_{x,y} p(x, y) \log p(x, y).$$

Using the factorization  $p(x, y) = p(x)p(y | x)$ ,

$$\begin{aligned} H(X, Y) &= - \sum_{x,y} p(x, y) \log (p(x)p(y | x)) \\ &= - \sum_{x,y} p(x, y) \log p(x) - \sum_{x,y} p(x, y) \log p(y | x). \end{aligned}$$

The first term simplifies to

$$- \sum_x p(x) \log p(x) = H(X),$$

and the second term is, by definition,  $H(Y | X)$ . Summing yields the desired identity.  $\square$

### (c) Non-negativity of KL Divergence

**Theorem.** For probability mass functions  $p$  and  $q$  on the same support,

$$\text{KL}(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \geq 0.$$

*Proof.* Rewrite the KL divergence as

$$\text{KL}(p||q) = \mathbb{E}_p \left[ -\log \frac{q(X)}{p(X)} \right].$$

The function  $f(t) = -\log t$  is convex. By Jensen's inequality,

$$\mathbb{E}_p[f(Z)] \geq f(\mathbb{E}_p[Z]),$$

where  $Z = \frac{q(X)}{p(X)}$ . Therefore,

$$\text{KL}(p||q) \geq -\log \left( \mathbb{E}_p \left[ \frac{q(X)}{p(X)} \right] \right).$$

But

$$\mathbb{E}_p \left[ \frac{q(X)}{p(X)} \right] = \sum_x p(x) \frac{q(x)}{p(x)} = \sum_x q(x) = 1.$$

Hence  $\text{KL}(p||q) \geq -\log 1 = 0$ .  $\square$

## (d) Mutual Information as KL Divergence

**Theorem.** *The mutual information satisfies*

$$I(X; Y) = \text{KL}(p(x, y) \parallel p(x)p(y)).$$

*Proof.* By definition,

$$I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}.$$

This expression is exactly the KL divergence between the joint distribution  $p(x, y)$  and the product of marginals  $p(x)p(y)$ . Hence,

$$I(X; Y) = \text{KL}(p(x, y) \parallel p(x)p(y)).$$

□

## Question 4: Experimental Results and Visualizations

In this question, we analyze the performance of different classifiers on the dataset and visualize the results.

The accuracy of the decision tree classifier on the validation set is: **[0.8121]**, with parameters `{'max_depth': None, 'criterion': 'gini'}`.

Decision Tree Hyperparameter Sweep:

```
=====
max_depth=2    , criterion=gini      -> Validation Accuracy: 0.7564
max_depth=2    , criterion=entropy   -> Validation Accuracy: 0.7564
max_depth=5    , criterion=gini      -> Validation Accuracy: 0.7707
max_depth=5    , criterion=entropy   -> Validation Accuracy: 0.7723
max_depth=10   , criterion=gini      -> Validation Accuracy: 0.7930
max_depth=10   , criterion=entropy   -> Validation Accuracy: 0.7739
max_depth=20   , criterion=gini      -> Validation Accuracy: 0.7882
max_depth=20   , criterion=entropy   -> Validation Accuracy: 0.7850
max_depth=None , criterion=gini      -> Validation Accuracy: 0.8121
max_depth=None , criterion=entropy   -> Validation Accuracy: 0.7930
=====
```

## (c) Decision Tree Visualization

Figure 3 shows the visualization of the trained decision tree.

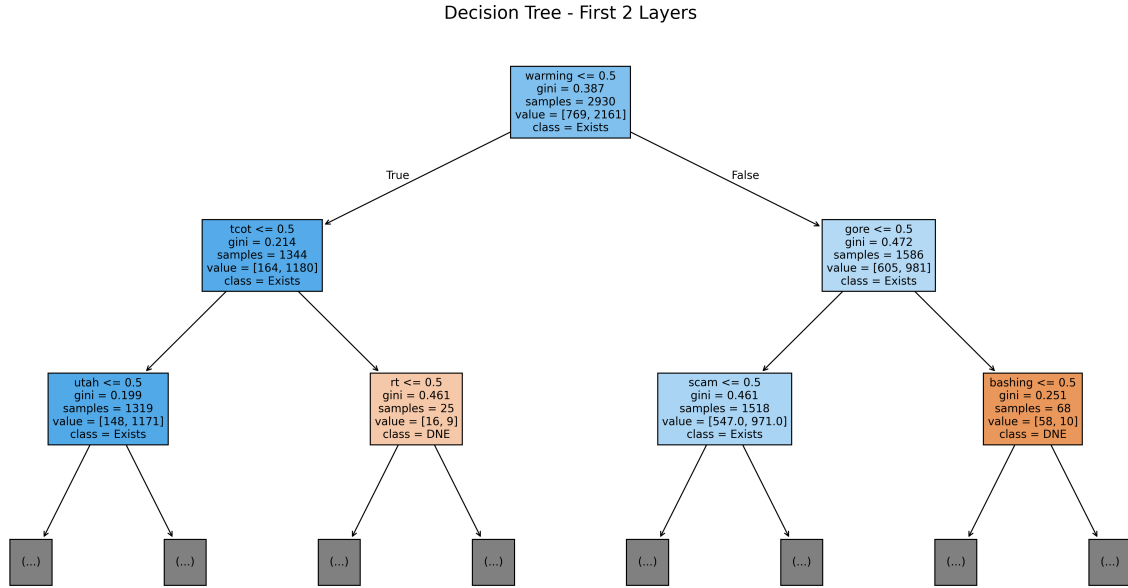


Figure 3: Visualization of the trained decision tree.

#### (d) Computing Information Gain

Information Gain for Top 10 Most Frequent Features:

=====	
Feature: http	Information Gain: 0.015375
Feature: global	Information Gain: 0.092796
Feature: warming	Information Gain: 0.091983
Feature: climate	Information Gain: 0.071898
Feature: change	Information Gain: 0.063709
Feature: bitly	Information Gain: 0.006933
Feature: rt	Information Gain: 0.016528
Feature: snow	Information Gain: 0.014682
Feature: new	Information Gain: 0.001755
Feature: retwtme	Information Gain: 0.005028

#### (e) KNN Error Plot

Figure 4 displays the k-nearest neighbors error as a function of the number of neighbors.

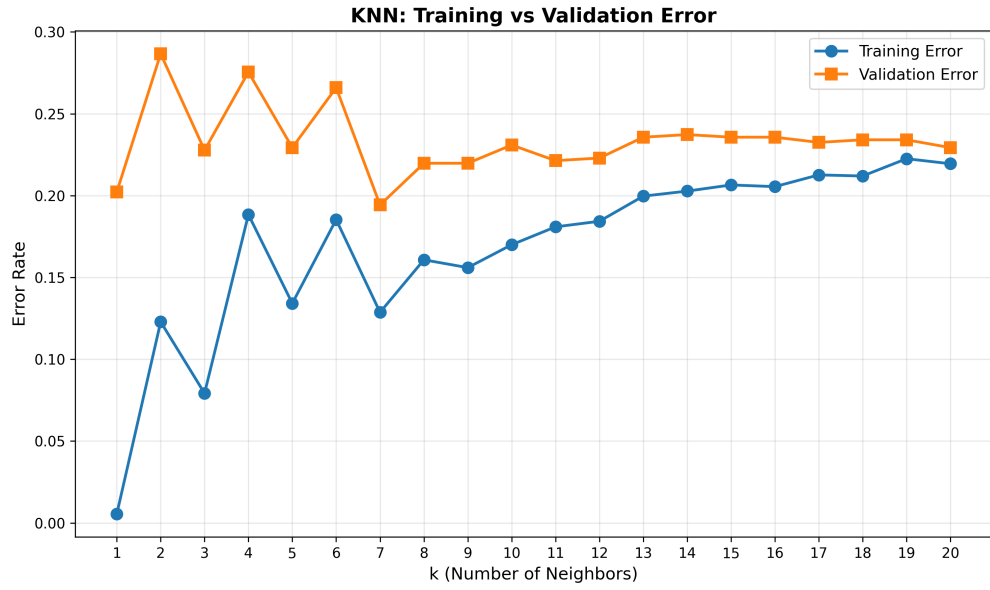


Figure 4: K-nearest neighbors error plot as a function of the number of neighbors.

Best k: 7

Training error at best k: 0.1287

Validation error at best k: 0.1943