MSDM-5054 homework-1

TIANYAO TAN

September 22 2025

1 Problem 1: Basic Knowledge

1.1 Question:

"Overfitting occurs only when training error is zero." Is the claim True or Wrong? Explain!

1.1.1 Answer:

This statement is incorrect.

Explanation:

Overfitting occurs when a model learns the training data (including noise) too well, resulting in poor performance on the test set. The key characteristic of overfitting is a significant gap between training error and validation error; specifically, excellent performance during training and poor performance during testing.

While a training error of zero is a clear symptom of overfitting, indicating that the model has perfectly memorized the training data, this does not necessarily mean that overfitting has occurred. For example, a model with a very small but non-zero training error is still severely overfitted if its test error is significantly above zero, indicating poor generalization. The core issue is that the model fails to generalize, performing poorly on the test set, but not necessarily achieving zero error on the training set.

1.2 Question:

When K increases, does the variance of KNN decrease or increase? Explain!

1.2.1 Answer:

- Variance measures how much a model's predictions change when trained on different datasets.

- In KNN, the parameter K determines how many neighbors influence the prediction.
- When K is small, predictions rely on very few points, making the model highly sensitive to noise in the training data → high variance.
- As K grows, predictions are based on a larger group of neighbors. This averaging effect smooths the decision boundary and reduces sensitivity to individual data points → lower variance.
- The tradeoff: while variance decreases, bias increases because the model becomes less flexible and may overlook local patterns.

1.3 Question:

Consider classification problem with the response variable Y taking possible values from $\{1, \cdots, K\}$ and denote $q_j(x) = \mathbb{P}(Y = j|X = x), \forall j = 1, \cdots, K$ the probability that $Y = j$ given $X = x$. What is the expected misclassification error of the *Bayes Classifier*? You can use $p(x)$ to represent the probability density function of X.

1.3.1 Answer:

Let:

- $Y \in \{1, 2, \ldots, K\}$ be the response variable (class label).
- $X \in \mathbb{R}^d$ be the feature vector.
- $q_j(x) = \mathbb{P}(Y = j \mid X = x)$ be the posterior probability of class j given $X = x$.
- $p(x)$ be the probability density function of X.

  The Bayes classifier:

$$\hat{y}(x) = \arg \max_{j \in \{1,\ldots,K\}} q_j(x)$$

Bayes Classifier and Misclassification Error:

$$\mathbb{E}[\mho\{\hat{Y}(X) \neq Y\}]$$

where $\mho$ is the indicator function.

The conditional misclassification error is:

$$\mathbb{P}(\hat{Y}(x) \neq Y \mid X = x) = 1 - \max_j q_j(x)$$

Bayes classifier will choose the class with the highest posterior probability.

Expected Error:

$$\mathbb{E}[\mho\{\hat{Y}(X) \neq Y\}] = \int \left(1 - \max_j q_j(x)\right) p(x)\, dx$$

$$\mathbb{E}[\mho\{\hat{Y}(X) \neq Y\}] = \int_{\mathbb{R}^d} \left(1 - \max_{j \in \{1,\ldots,K\}} \mathbb{P}(Y = j \mid X = x)\right) p(x)\, dx$$

1.4 Question:

Show that the sum of residuals in simple linear regression is zero. How about in multiple linear regression?

1.4.1 Answer:

In simple linear regression, we have a model of the form:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

The residual, e_i, for each observation i is the difference between the observed value (Y_i) and the predicted value ($\hat{Y}_i$):

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

We can look at the first-order conditions for minimizing the SSR. The SSR is given by:

$$SSR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

we can take the partial derivatives with respect to each and set them to zero.

The partial derivative with respect to $\hat{\beta}_0$:

$$\frac{\partial SSR}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

Setting this to zero to find the minimum:

$$-2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

Dividing by -2, we get:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i) = 0$$

Since $e_i = Y_i - \hat{Y}_i$

Simplify to:

$$\sum_{i=1}^{n} e_i = 0$$

Thus, the sum of the residuals is zero.

The multiple linear regression includes multiple independent variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$$

The predicted value is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_p X_p$.

The residuals are still defined as $e_i = Y_i - \hat{Y}_i$.

We also minimize the sum of squared residuals:

$$SSR = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - ... - \hat{\beta}_p X_{ip})^2$$

When we take the partial derivative with respect to the intercept term, $\hat{\beta}_0$, and set it to zero, we arrive at the same result:

$$\frac{\partial SSR}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - ... - \hat{\beta}_p X_{ip}) = 0$$

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i) = 0$$

$$\sum_{i=1}^{n} e_i = 0$$

Therefore, the sum of the residuals in multiple linear regression is also zero.

1.5 Question:

Given two linear models with the same number of parameters, can we use R square to compare the performance of the two models, or must we use the adjusted R square? Explain!

1.5.1 Answer:

We can use R-Squared to compare them.

The model with the higher R-squared value provides a better fit to the data.

R^2 tells us how well our model's predictions approximate the real data points.

The adjusted R-squared incorporates a penalty for each additional variable included.

Since two linear models have an identical number of predictors, the penalty applied by the adjusted R-squared calculation would be the same for both. Consequently, the ranking of the two models using R-squared would be the same as the ranking using adjusted R-squared. The model with the higher R-squared will also have the higher adjusted R-squared.

Therefore, R square provides a direct comparison of the proportion of variance explained without needing adjustment for differing model complexities.

1.6 Question:

Explain the advantage and disadvantage of linear regression and KNN regression.

1.6.1 Answer:

Advantages of Linear Regression:

The model's output is a straightforward equation where the coefficients represent the strength and direction of the relationship between each feature and the target.

Linear regression is very fast to train, especially on large datasets. The mathematical solution to find the best-fit line is well-established and computationally inexpensive.

Because it makes a strong assumption about the data's structure, it tends to have low variance. This stability can be beneficial in preventing overfitting.

Disadvantages of Linear Regression:

If the true relationship between the features and the target is not linear, the model will have a high bias and will not perform well.

Outliers can have a significant impact on the position of the regression line and skew the model's predictions for the rest of the data.

Advantages of KNN Regression:

Since KNN makes no assumptions about the data, it can capture complex, non-linear relationships. The prediction is based on the average of the target values of its k nearest neighbors, allowing it to adapt to the local structure of the data.

Disadvantages of KNN Regression:

With a large dataset, the algorithm has to calculate the distance to every single point in the training data to find the nearest neighbors.

KNN's performance can degrade significantly as the number of features increases. In high-dimensional spaces, the concept of distance becomes less meaningful.

KNN relies on distance metrics, the feature with the larger scale will dominate the distance calculation.