CS129: Applied Machine Learning
Discussion Section 5

1 Adding a Predictor in Logistic Regression

Suppose we have a classification problem with a binary response Y and a p-dimensional predictor vector $X = (X_1, \ldots, X_p)$. Logistic regression is fit to a set of n samples. Then, logistic regression is fit again to the same observations, but now we include one additional predictor, so that

$$X = (X_1, \ldots, X_p, X_{p+1}).$$

Explain how the **training error**, **test error**, and **coefficients** change in each of the following cases:

(a) $X_{p+1} = X_1 + 2X_p$.

(b) X_{p+1} is a random variable independent of Y.

2 Multi-Task Learning vs. Separate Models

You are given a dataset of patients who visited UPMC Hospital in 2011. For each patient, a set of features (e.g., temperature, height, lab results) has been extracted.

Your goal is to predict whether a new patient has any of the following conditions:

- Diabetes
- Heart disease
- Alzheimer's disease

A patient may have zero, one, or multiple conditions.

You are considering two modeling approaches:

Approach A: Train three separate neural networks — one for each disease.

Approach B: Train a single neural network with three output neurons (one per disease) that share a common hidden representation.

Discussion Questions

Which method do you prefer? Briefly justify your answer.

3 When Validation Performance Misleads You

You split your dataset into three subsets: training, validation (dev), and test.

You train 10 different models on the training set and select the model that achieves the lowest error on the validation set.

The selected model achieves strong performance on the validation set. However, when evaluated on the test set, its error is significantly worse.

(a) Give two possible explanations for this behavior.

(b) For each explanation, describe briefly what you would check or change to diagnose the issue.

4 Estimating μ Under Noise

Assume we want to estimate an unknown quantity μ. We observe i.i.d. samples

$$Y_1, Y_2, \ldots, Y_n,$$

where each $Y_i \sim \mathcal{N}(\mu, \sigma^2)$.

(a) How would you estimate μ from the samples?

(b) The estimator in (a) can be sensitive when σ is large. What can you do to reduce the impact of noise?

(c) If sampling is expensive, suggest a different estimator for μ. How does its bias/variance compare to the sample mean?

5 Bias–Variance Tradeoff (Conceptual Questions)

(a) Describe a setting in which **bias** is worse than variance. Briefly justify your answer.

(b) Describe a setting in which **variance** is worse than bias. Briefly justify your answer.

6 Bias, Variance, and Model Flexibility

In the context of bias-variance tradeoff, flexibility is the capacity of a model class to fit a wide variety of functions. More formally:

- A more flexible model can represent more complex patterns in the data.
- A less flexible model is more constrained in the shapes of functions it can represent.

(i) Suppose you have regression data generated by a polynomial of degree 3. Characterize the bias-variance of the estimates of the following models on the data with respect to the true model by circling the appropriate entry.

Model	Bias	Variance
Linear regression	low / high	low / high
Polynomial regression with degree 3	low / high	low / high
Polynomial regression with degree 10	low / high	low / high

(ii) For each part below indicate whether we would generally expect the performance of a **flexible** statistical learning method to be **better** or **worse** than an **inflexible** method. Justify your answer.

(i) The sample size n is extremely large, and the number of predictors p is small.

(ii) The number of predictors p is extremely large, and the number of observations n is small.

(iii) The relationship between the predictors and response is highly non-linear.

(iv) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\varepsilon)$, is extremely high.

7 Bias–Variance from Learning Curves

The following graphs plots the error distribution of three models. Comment on each model in terms of bias-variance.

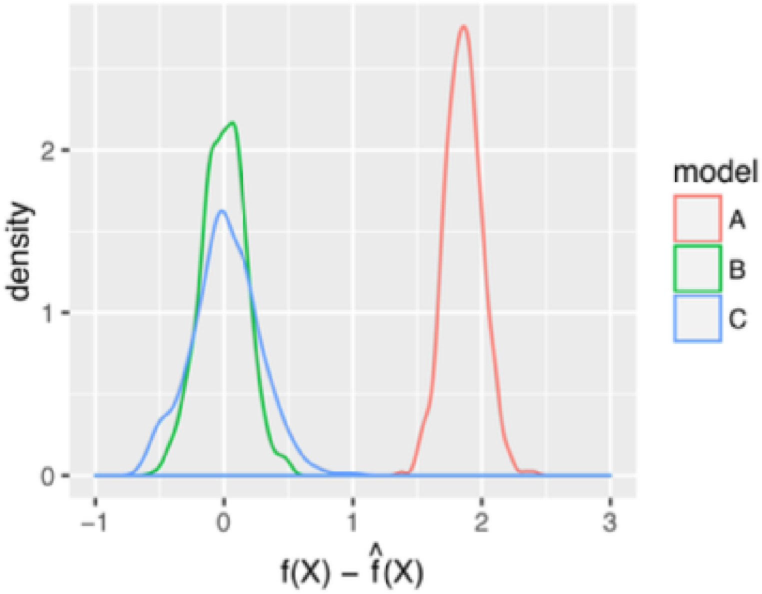

Figure 1: Learning curves for three different models.

The three figures above show the difference between bias-variance in terms of prediction.

8 Diagnosing Gradient Descent Behavior

The following plots show the cost function $J(\theta)$ versus the number of iterations for three different learning rates.

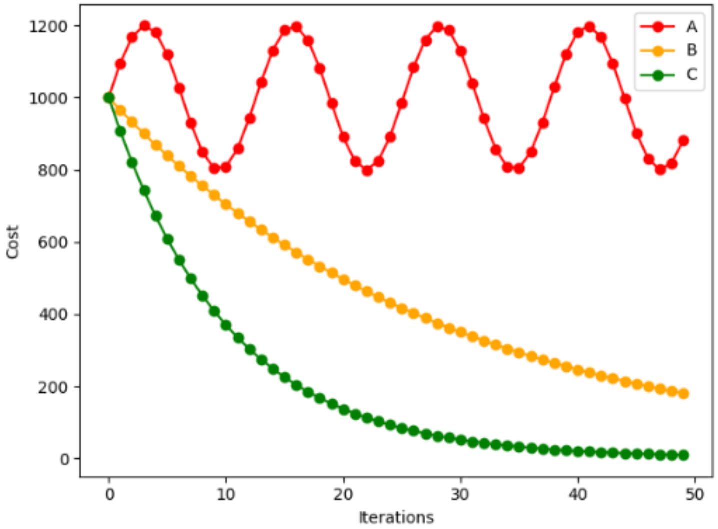

Figure 2: Cost vs. iterations for three different learning rates.

(a) Which plot corresponds to a learning rate that is too high? Justify your answer.

(b) Which plot corresponds to a learning rate that is too low? What would you adjust?

(c) In Plot C, the cost flattens out after some iterations. What does this indicate?

(d) If the cost never meaningfully decreases despite many iterations, what might be wrong?