

Exercise Sheet 10

Exercise 1: Bias and Variance of Mean Estimators (20 P)

Assume we have an estimator $\hat{\theta}$ for a parameter θ . The bias of the estimator $\hat{\theta}$ is the difference between the true value for the estimator, and its expected value:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta} - \theta].$$

If $\text{Bias}(\hat{\theta}) = 0$, then $\hat{\theta}$ is called unbiased. The variance of the estimator $\hat{\theta}$ is the expected square deviation from its expected value:

$$\text{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2].$$

The mean squared error of the estimator $\hat{\theta}$ is

$$\text{Error}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$$

Let X_1, \dots, X_N be a sample of i.i.d random variables. Assume that X_i has mean μ and variance σ^2 . Calculate the bias, variance and mean squared error of the following mean estimators:

- (a) $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N X_i$ (i.e. the sample mean),
- (b) $\hat{\mu} = 0$.

Exercise 2: Bias-Variance Decomposition for Regression (15 P)

Let $y = f(x)$ be a function mapping input to output and evaluated at some out-of-sample data point x . Consider an estimator $\hat{f}(x)$ that is obtained by training a regression model on some random sample $\mathcal{D} = \{(x_1, y_1), \dots, (x_N, y_N)\}$ of the function $y = f(x)$.

- (a) Prove the bias-variance decomposition

$$\text{Error}(\hat{f}(x)) = \text{Bias}(\hat{f}(x))^2 + \text{Var}(\hat{f}(x))$$

where the mean squared error, bias and variance are given by

$$\text{Error}(\hat{f}(x)) = \mathbb{E}[(\hat{f}(x) - f(x))^2], \quad \text{Bias}(\hat{f}(x)) = \mathbb{E}[\hat{f}(x) - f(x)], \quad \text{Var}(\hat{f}(x)) = \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2].$$

Exercise 3: Bias-Variance Decomposition for Classification (15 P)

The bias-variance decomposition usually applies to regression data. In this exercise, we would like to obtain similar decomposition for classification, in particular, when the prediction is given as a probability distribution over C classes. Let $P = [P_1, \dots, P_C]$ be the ground truth class distribution associated to a particular input pattern. Assume a random estimator of class probabilities $\hat{P} = [\hat{P}_1, \dots, \hat{P}_C]$ for the same input pattern. The error function is given by the KL-divergence between the ground truth and the estimated probability distribution:

$$\text{Error} = \mathbb{E}[D_{\text{KL}}(P || \hat{P})].$$

First, we would like to determine the mean of the class distribution estimator \hat{P} . We define the mean as the distribution that minimizes its expected KL divergence from the the class distribution estimator, that is, the distribution R that optimizes

$$\min_R \mathbb{E}[D_{\text{KL}}(R || \hat{P})].$$

- (a) Show that the solution to the optimization problem above is given by

$$R = [R_1, \dots, R_C] \quad \text{where} \quad R_i = \frac{\exp \mathbb{E}[\log \hat{P}_i]}{\sum_j \exp \mathbb{E}[\log \hat{P}_j]} \quad \forall 1 \leq i \leq C.$$

- (b) Prove the bias-variance decomposition

$$\text{Error}(\hat{P}) = \text{Bias}(\hat{P}) + \text{Var}(\hat{P})$$

where the error, bias and variance are given by

$$\text{Error}(\hat{P}) = \mathbb{E}[D_{\text{KL}}(P || \hat{P})], \quad \text{Bias}(\hat{P}) = D_{\text{KL}}(P || R), \quad \text{Var}(\hat{P}) = \mathbb{E}[D_{\text{KL}}(R || \hat{P})].$$

Exercise 4: Programming Exercise (50 P)

Download the programming files for exercise sheet 10 on ISIS and follow the instructions.