Fachgebiet Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 7

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

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

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

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

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

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

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 mean estimator:

$$\hat{\mu} = \alpha \cdot \frac{1}{N} \sum_{i=1}^{N} X_i$$

where α is a parameter between 0 and 1.

Exercise 2: Bias-Variance Decomposition for Classification (30 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 expected KL-divergence between the ground truth and the estimated probability distribution:

$$\text{Error} = \text{E}\big[D_{\text{KL}}(P||\hat{P})\big] = \text{E}\big[\textstyle\sum_{i=1}^{C} P_i \log(P_i/\hat{P}_i)\big].$$

First, we would like to determine the mean of 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

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

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

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

(Hint: To implement the positivity constraint on R, you can reparameterize its components as $R_i = \exp(Z_i)$, and minimize the objective w.r.t. Z.)

(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

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

(Hint: as a first step, it can be useful to show that $\text{E}[\log R_i - \log \hat{P}_i]$ does not depend on the index i.)

Exercise 3: Programming (50 P)

Download the programming files on ISIS and follow the instructions.