Exercises for the course

Machine Learning 1

Winter semester 2021/22

Abteilung 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 5

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

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

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

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

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

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

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:

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

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

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

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

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

(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

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

where the error, bias and variance are given by

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

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

Exercise 3: Programming (50 P)

Download the programming files on ISIS and follow the instructions.

ML1 Sheet 5

1

a) The bias is defined as $\text{Bias}(\hat{\mu}) = E[\hat{\mu} - \mu]$

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

2) $E[\frac{\alpha}{N}] = \frac{\alpha}{N}$ and $E[\mu] = \mu$

3) $\mu = \frac{1}{N} \sum_{i=1}^{N} E[X_i]$

$$\overset{1)}{=} E[\frac{\alpha}{N} \sum_{i=1}^{N} X_i - \mu]$$

$$\overset{2)}{=} \frac{\alpha}{N} \sum_{i=1}^{N} E[X_i] - \mu$$

$$\overset{3)}{=} \alpha\mu - \mu = (\alpha - 1)\mu$$

b) $\text{Var}(\hat{\mu}) = \text{Var}(\frac{\alpha}{N} \sum_{i=1}^{N} X_i) = (\frac{\alpha}{N})^2 \sum_{i=1}^{N} \text{Var}(X_i)$

$$= \frac{\alpha^2}{N^2} \sum_{i=1}^{N} \sigma^2 = \frac{\alpha^2}{N^2} \sigma^2$$

c) $\text{Error}(\hat{\mu}) = (\text{Bias}(\hat{\mu}))^2 + \text{Var}(\hat{\mu})$

$$= \mu^2 (\alpha - 1)^2 + (\frac{\alpha}{N} \sigma)^2$$

2

a) We need to solve $\min_R E[D_{KL}(R \| \hat{P})]$.

$$\min_R E[D_{KL}(R\|\hat{P})] = \min_R E[\sum_{i=1}^{c} R_i \log(R_i/\hat{P}_i)$$

$$= \min_R E[\sum_{i=1}^{c} R_i \log R_i - R_i \log \hat{P}_i]$$

$$= \min_R \sum_{i=1}^{c} R_i \log R_i - R_i E[\log \hat{P}_i]$$

$$= \min_R \sum_{i=1}^{c} e^{z_i} z_i - e^{z_i} E[\log \hat{P}_i] \quad \text{using } R_i = e^{z_i}$$

Using the contraint $\sum_{i=1}^{c} e^{z_i} = 1$ we can set the Lagrangian:

$$\mathcal{L}(z,\lambda) = \sum_{i=1}^{c} e^{z_i} z_i - e^{z_i} E[\log \hat{P}_i] + \lambda(\sum_{i=1}^{c} e^{z_i} - 1)$$

$$\frac{\partial \mathcal{L}}{\partial z} = \sum_{i=1}^{c} e^{z_i} + e^{z_i} z_i - e^{z_i} E[\log \hat{P}_i] + e^{z_i} \lambda$$

$$= \sum_{i=1}^{c} e^{z_i}(1 + z_i - E[\log \hat{P}_i] + \lambda) \overset{!}{=} 0$$

In this product $p \cdot q = 0$, $p = e^{z_i}$ can never be zero, hence the other term must become 0:

$$1 + z_i - E[\log \hat{P}_i] + \lambda \overset{!}{=} 0 \Leftrightarrow z_i = E[\log \hat{P}_i] - 1 - \lambda$$

$$\Leftrightarrow R_i = e^{E[\log \hat{P}_i] - 1 - \lambda} = \exp(E[\log \hat{P}_i]) / \exp(1+\lambda)$$

This becomes $R_i = \dfrac{\exp E[\log \hat{P}_i]}{\sum_j \exp E[\log \hat{P}_j]}$

using the remaining contraint $\exp(1+\lambda) = \sum_{i=1}^{c} \exp(E[\log \hat{P}_i])$

b) Using $R_i = \exp(E[\log \hat{P}_i])/\exp(1+\lambda)$ we write

$$E[\log R_i - \log \hat{P}_i] = E[\log(\exp(E[\log \hat{P}_i]))$$
$$- \log(\exp(1+\lambda)) - \log \hat{P}_i]$$

$$= E[E[\log \hat{P}_i] - (1+\lambda) - \log \hat{P}_i]$$
$$= E[\log \hat{P}_i] - (1+\lambda) - E[\log \hat{P}_i]$$
$$= -1-\lambda$$ which is independent of the index i.

The error is given by

$$\mathrm{Error}(\hat{P}) = E[D_{KL}(P \| \hat{P})] = E[\sum_{i=1}^{c} P_i \log(P_i/\hat{P}_i)]$$
$$= E[\sum_{i=1}^{c} P_i \log P_i - P_i \log \hat{P}_i] \cancel{= E[\sum_{i=1} P_i \log P_i + P_i \log}$$
$$= E[\underbrace{\sum_{i=1}^{c} P_i \log P_i - P_i \log R_i}_{\mathrm{Bias}(\hat{P})} + P_i \log R_i - P_i \log \hat{P}_i]$$
$$= \mathrm{Bias}(\hat{P}) + E[\sum_{i=1}^{c} P_i \log R_i - P_i \log \hat{P}_i]$$
$$= \mathrm{Bias}(\hat{P}) + \sum_{i=1}^{c} P_i\, E[\log R_i - \log \hat{P}_i]$$

Since $\sum_{i=1}^{c} R_i = \sum_{i=1}^{c} P_i = 1$

$$= \mathrm{Bias}(\hat{P}) + \sum_{i=1}^{c} R_i\, E[\log R_i - \log \hat{P}_i]$$
$$= \mathrm{Bias}(\hat{P}) + E[\sum_{i=1}^{c} R_i \log R_i - R_i \log \hat{P}_i]$$

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