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 $\theta$. 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 $\mu$ and variance $\sigma^2$. *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 $\alpha$ 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.

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

Assume we have an estimator $\hat{\theta}$ for a parameter $\theta$. 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 $\mu$ and variance $\sigma^2$. *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 $\alpha$ is a parameter between 0 and 1.

$$\begin{aligned}
\mathrm{Bias}(\hat{\mu}) = \mathbb{E}[\hat{\mu} - \mu] &= \mathbb{E}\Big[\alpha \cdot \frac{1}{N}\sum_{i=1}^{N} X_i - \mu\Big] \\
&= \frac{\alpha}{N}\sum_{i=1}^{N} \mathbb{E}[X_i] - \mu \\
&= \frac{\alpha}{N}\sum_{i=1}^{N} \mu - \mu \\
&= \frac{\alpha}{\cancel{N}} \cdot \cancel{N}\mu - \mu \\
&= (\alpha - 1)\mu
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}[\hat{\mu}] = \mathbb{V}\Big[\frac{\alpha}{N}\sum_{i=1}^{N} X_i\Big] &= \Big(\frac{\alpha}{N}\Big)^2 \cdot \sum_{i=1}^{N} \mathbb{V}[X_i] \\
&= \Big(\frac{\alpha}{N}\Big)^2 \cdot \sum_{i=1}^{N} \sigma^2 \\
&= \frac{\alpha^2}{N^2} \cdot N \cdot \sigma^2 \\
&= \frac{\alpha^2}{N} \cdot \sigma^2
\end{aligned}$$

$$\mathrm{Error}(\hat{\mu}) = \mathrm{Bias}(\hat{\mu})^2 + \mathbb{V}[\hat{\mu}] = (\alpha - 1)^2 \mu^2 + \frac{\alpha^2}{N} \cdot \sigma^2$$

**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:

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

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

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

*(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. $Z$.)*

$$\min_R \mathbb{E}[D_{KL}(R||\hat{P})] = \min_R \mathbb{E}\Big[\sum_{i=1}^{C} R_i \cdot \log\Big(\frac{R_i}{\hat{P}_i}\Big)\Big]$$

$$= \min_R \mathbb{E}\Big[\sum_{i=1}^{C} R_i \log(R_i) - \sum_{i=1}^{C} R_i \log(\hat{P}_i)\Big]$$

$$= \min_R \sum_{i=1}^{C} R_i \log(R_i) - \sum_{i=1}^{C} R_i\, \mathbb{E}[\log(\hat{P}_i)]$$

这里E是对estimater求mean

(To make sure $R_i > 0 \ \forall\, 1 \le i \le C$, We can reparameterize it as $R_i = e^{z_i}$)

$$= \min_R \sum_{i=1}^{C} e^{z_i} \log(e^{z_i}) - \sum_{i=1}^{C} e^{z_i}\, \mathbb{E}[\log(\hat{P}_i)]$$

$$= \min_R \sum_{i=1}^{C} z_i \cdot e^{z_i} - \sum_{i=1}^{C} e^{z_i}\, \mathbb{E}[\log(\hat{P}_i)]$$

subject to $\sum_{i=1}^{C} R_i = 1 = \sum_{i=1}^{C} e^{z_i}$

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

$$\frac{\partial}{\partial z_i} \mathcal{L}(z,\lambda) = e^{z_i} + z_i e^{z_i} - e^{z_i}\, \mathbb{E}[\log(\hat{P}_i)] + \lambda \cdot e^{z_i}$$

$$= e^{z_i} \cdot \big(1 + z_i - \mathbb{E}[\log(\hat{P}_i)] + \lambda\big) = 0$$

$$\Rightarrow z_i = \mathbb{E}[\log(\hat{P}_i)] - 1 - \lambda \qquad ①$$

$$\frac{\partial}{\partial \lambda} \mathcal{L}(z,\lambda) = \sum_{i=1}^{C} e^{z_i} - 1 = \sum_{i=1}^{C} e^{\mathbb{E}[\log(\hat{P}_i)] - 1 - \lambda} - 1 = 0$$

$$\sum_{i=1}^{C} e^{\mathbb{E}[\log(\hat{P}_i)]} = e^{1+\lambda}$$

$$1 + \lambda = \log\Big[\sum_{i=1}^{C} e^{\mathbb{E}[\log(\hat{P}_i)]}\Big] \qquad ②$$

from ①,②: $z_i = \mathbb{E}[\log(\hat{p}_i)] - 1 - \lambda$

$$= \mathbb{E}[\log(\hat{p}_i)] - \log\left[\sum_{i=1}^{c} e^{\mathbb{E}[\log(\hat{p}_i)]}\right]$$

$$R_i = e^{z_i} = e^{\mathbb{E}[\log(\hat{p}_i)] - \log\left[\sum_{i=1}^{c} e^{\mathbb{E}[\log(\hat{p}_i)]}\right]}$$

$$= \frac{e^{\mathbb{E}[\log(\hat{p}_i)]}}{\sum_{i=1}^{c} e^{\mathbb{E}[\log(\hat{p}_i)]}}$$

(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* $i$*.)*

$$\mathbb{E}[\log R_i - \log \hat{P}_i] = \mathbb{E}\left[\log\left(\frac{e^{\mathbb{E}[\log(\hat{P}_i)]}}{\sum_{i=1}^{c} e^{\mathbb{E}[\log(\hat{P}_i)]}}\right) - \log \hat{P}_i\right]$$

$$= \mathbb{E}\left[\mathbb{E}[\log(\hat{P}_i)] - \log\left(\sum_{i=1}^{c}\left[e^{\mathbb{E}[\log(\hat{P}_i)]}\right]\right) - \log \hat{P}_i\right]$$

$$= \underbrace{\mathbb{E}\left[\mathbb{E}[\log(\hat{P}_i)] - \log \hat{P}_i\right]}_{=0} - \mathbb{E}\left[\log\left(\sum_{i=1}^{c}\left[e^{\mathbb{E}[\log(\hat{P}_i)]}\right]\right)\right]$$

$$= -\mathbb{E}\left[\log\left(\sum_{i=1}^{c}\left[e^{\mathbb{E}[\log(\hat{P}_i)]}\right]\right)\right]$$

from 2.a) ②: $= -\mathbb{E}[1+\lambda] = -1-\lambda$ which is independent of index $i$

$$\mathrm{Error}(\hat{P}) = \mathbb{E}[D_{KL}(P||\hat{P})] = \mathbb{E}\left[\sum_{i=1}^{c} P_i \log\left(\frac{P_i}{\hat{P}_i}\right)\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{c} P_i \log P_i - \sum_{i=1}^{c} P_i \log \hat{P}_i\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{c} P_i \log P_i - \sum_{i=1}^{c} P_i \log R_i + \sum_{i=1}^{c} P_i \log R_i - \sum_{i=1}^{c} P_i \log \hat{P}_i\right]$$

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

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

$$= \mathrm{Bias}(\hat{P}) + \sum_{i=1}^{c}\left[P_i \cdot (-1-\lambda)\right]$$

$$= \mathrm{Bias}(\hat{P}) + (-1-\lambda)\sum_{i=1}^{c} P_i$$ Because of $\sum_{i=1}^{c} P_i = \sum_{i=1}^{c} R_i = 1$

$$= \mathrm{Bias}(\hat{P}) + (-1-\lambda)\sum_{i=1}^{c} R_i$$

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

$$= \mathrm{Bias}(\hat{P}) + \mathbb{E}\left[\sum_{i=1}^{c} R_i \cdot \log \frac{R_i}{\hat{P}_i}\right]$$

$$= \mathrm{Bias}(\hat{P}) + \mathbb{V}[\hat{P}]$$