Exercises for the course
Machine Learning 1
Winter semester 2023/24

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

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

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

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

(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.)

(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 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 $X_1, \ldots, X_N$ 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.

Solution:

(1) Bias:

$$\text{Bias}(\hat{\mu}) = E[\hat{\mu} - \mu] = E\left[\alpha \cdot \frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right]$$
$$= \alpha\, E\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] - \mu$$
$$= (\alpha - 1)\mu$$

(2) Variance:

$$\text{Var}(\hat{\mu}) = E[(\hat{\mu} - E[\hat{\mu}])^2]$$
$$= \text{Var}\left(\alpha \cdot \frac{1}{N}\sum_{i=1}^{N} X_i\right)$$
$$= \frac{\alpha^2}{N^2}\text{Var}\left(\sum_{i=1}^{N} X_i\right)$$

Since $X_1, X_2 \cdots, X_N$ are i.i.d variables

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

(3) Error

$$\text{Error}(\hat{\mu}) = \text{Bias}^2(\hat{\mu}) + \text{Var}(\hat{\mu})$$
$$= (\alpha-1)^2\mu^2 + \frac{\alpha^2}{N}\sigma^2$$

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.)

Solution:

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

If we replace the R_i with $R_i = \exp(Z_i)$ and consider the fact that R_i is a probability, then we can reformulate the optimization

$$\boxed{\begin{aligned} &\min_Z \mathrm{E}\left[\sum_{i=1}^{C} \exp(Z_i)\left(\log(\exp(Z_i)) - \log \hat{P}_i\right)\right] \\ &\text{s.t.} \quad \sum_{i=1}^{C} \exp(Z_i) = 1 \end{aligned}}$$

$$\Downarrow$$

$$\boxed{\begin{aligned} &\min_Z \mathrm{E}\left[\sum_{i=1}^{C} \exp(Z_i) Z_i - \exp(Z_i)\log \hat{P}_i\right] \\ &\text{s.t.} \quad \sum_{i=1}^{C} \exp(Z_i) = 1 \end{aligned}}$$

日期: /

$$\boxed{\begin{aligned} &\min_{Z} \sum_{i=1}^{C} \exp(Z_i) Z_i - \exp(Z_i) E[\log \hat{P}_i] \\ &\text{s.t.} \quad \sum_{i=1}^{C} \exp(Z_i) = 1 \end{aligned}}$$

Then we can use the lagrange multiplier to solve this constrainted optimization problem.

$$\mathcal{L}(Z, \lambda) = \sum_{i=1}^{C} \exp(Z_i) Z_i - \exp(Z_i) E[\log \hat{P}_i] + \lambda \left(\sum_i \exp(Z_i) - 1 \right)$$

compute the derivatives w.r.t Z_i and λ

$$\frac{\partial \mathcal{L}}{\partial Z_i} = \exp(Z_i)(Z_i + 1) - \exp(Z_i) E[\log \hat{P}_i] + \lambda \exp(Z_i)$$

$$= \exp(Z_i)(Z_i + 1 + \lambda) - \exp(Z_i) E[\log \hat{P}_i] = 0$$

Since $\exp(Z_i) > 0$, then:

$$Z_i + 1 + \lambda = E[\log \hat{P}_i] \longrightarrow Z_i = E[\log \hat{P}_i] - 1 - \lambda$$

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

Since R_i is a probability density, then we have

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

日期: /

which means that :

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

$$\therefore \quad R_i = \frac{\exp(E[\log \hat{P}_i])}{\sum_{i=1}^{C} \exp(E[\log \hat{P}_i])} \quad \forall i = 1 \cdots C$$

proofed.

(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*.)*

Solution:

From (a) we have :

$$R_i = \frac{\exp(E[\log \hat{P}_i])}{\sum_{i=1}^{C} \exp(E[\log \hat{P}_i])}$$

① First proof $E[\log R_i - \log \hat{P}_i]$ doesn't depend on index i.

$$E[\log R_i - \log \hat{P}_i] = E\Big[E[\log \hat{P}_i] - \log\big(\underbrace{\textstyle\sum_{i=1}^{C} \exp(E[\log \hat{P}_i])}_{M}\big) - \log \hat{P}_i\Big]$$

$$= E[\log \hat{P}_i] - E[\log \hat{P}_i] - E[\log M]$$

$$= -E_{(R_i)}[\log M]$$

$$= -\log M$$

日期:　　/

Since M is the sum of $i \in \{1, \cdots, C\}$

$\therefore$ $E[\log R_i - \log \hat{P}_i]$ is independent of the index.

② Then prove the bias-variance decomposition.

$$\begin{aligned}
\text{Error}(\hat{P}) &= E[D_{KL}(P \| \hat{P})] \\
&= E\left[\sum_i P_i \log P_i - P_i \log \hat{P}_i + P_i \log R_i - P_i \log R_i\right] \\
&= E\left[\sum_i P_i \log P_i - P_i \log R_i\right] + E\left[\sum_i -P_i \log \hat{P}_i + P_i \log R_i\right] \\
&= D_{KL}(P \| R) + \sum_i P_i \overbrace{E[\log R_i - \log \hat{P}_i]}^{\text{independent of index } i} \\
&= D_{KL}(P \| R) + E[\log R_i - \log \hat{P}_i] \cdot \left(\sum_i P_i\right)
\end{aligned}$$

Since P_i is a density function as the same as R_i

$\therefore$ $\sum_i P_i = \sum_i R_i = 1$

$$\begin{aligned}
&= D_{KL}(P \| R) + E[\log R_i - \log \hat{P}_i] \cdot \left(\sum_i R_i\right) \\
&= D_{KL}(P \| R) + E\left[\sum_i R_i \log R_i - R_i \log \hat{P}_i\right] \\
&= D_{KL}(P \| R) + E\left[D_{KL}(R \| \hat{P})\right] \\
&= \text{Bias}(\hat{P}) + \text{Var}(\hat{P})
\end{aligned}$$

proofed