UE 7

Sonntag, 10. November 2024 17:59

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 $X_1, \ldots, X_N$ 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.

(a) Bias

$$\begin{aligned}
\mathrm{Bias}(\hat{\mu}) &= \mathbb{E}(\hat{\mu} - \mu) && \text{linearity of } \mathbb{E} \\
&= \mathbb{E}(\hat{\mu}) - \mathbb{E}(\mu) && \text{Def } \hat{\mu},\ \mu \text{ const.} \\
&= \mathbb{E}\left(\alpha \tfrac{1}{N} \textstyle\sum_{i=1}^{N} X_i\right) - \mu && \text{linearity of } \mathbb{E} \\
&= \alpha \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \mathbb{E}(X_i) - \mu && \mathbb{E}(X_i) = \mu \\
&= \left(\alpha \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \mu\right) - \mu \\
&= \alpha \tfrac{1}{N} N \cdot \mu - \mu \\
&= \alpha \mu - \mu \\
&= (\alpha - 1)\mu
\end{aligned}$$

(b) Variance

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}) &= \mathrm{Var}\left(\alpha \tfrac{1}{N} \textstyle\sum_{i=1}^{N} X_i\right) && \mathrm{Var}(aZ) = a^2 \mathrm{Var}(Z) \\
&= \alpha^2 \tfrac{1}{N^2} \mathrm{Var}\left(\textstyle\sum_{i=1}^{N} X_i\right) && \forall i \neq j.\ X_i \perp X_j,\ X_i \text{ iid} \\
&= \alpha^2 \tfrac{1}{N^2} \textstyle\sum_{i=1}^{N} \mathrm{Var}(X_i) && \mathrm{Var}(X_i) = \sigma^2 \\
&= \alpha^2 \tfrac{1}{N^2} \textstyle\sum_{i=1}^{N} \sigma^2 \\
&= \alpha^2 \tfrac{N}{N^2} \sigma^2 \\
&= \alpha^2 \tfrac{1}{N} \sigma^2
\end{aligned}$$

(c) Error

$$\begin{aligned}
\mathrm{Error}(\hat{\mu}) &= \mathrm{Bias}(\hat{\mu})^2 + \mathrm{Var}(\hat{\mu}) && \text{Bias-Variance decomposition} \\
&= (\alpha - 1)^2 \mu^2 + \frac{\alpha^2}{N} \sigma^2
\end{aligned}$$

This estimator shrinks the estimation towards 0. In some cases, e.g. for centered data with $\mu = 0$, small values of α lead to a more accurate estimation of the mean with smaller Error than the sample average.

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

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

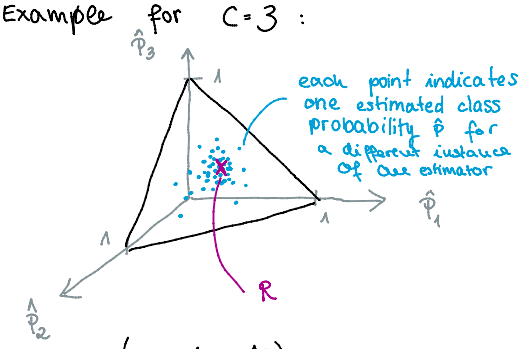
(a) *Show* that the solution to the optimization problem above is given by

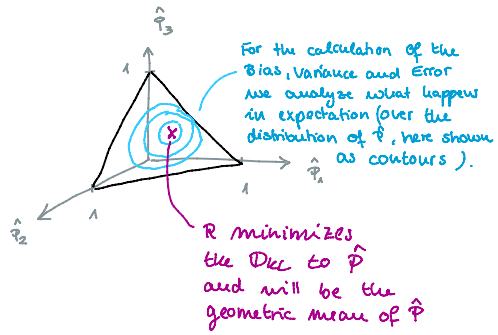
$$R = [R_1, \dots, R_C] \quad \text{where} \quad R_i = \frac{\exp \mathrm{E}\big[\log \hat{P}_i\big]}{\sum_j \exp \mathrm{E}\big[\log \hat{P}_j\big]} \qquad \forall\, 1 \le i \le C.$$

(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.)

E (expectation) is taken over the distribution of $\hat{P}_i$:

Example for $C=3$:





$$\min_z \ \mathbb{E}\big(D_{KL}(R \| \hat{P})\big) \qquad \text{(expectation over distribution of } \hat{P}\text{)}$$

$$= \min_z \ \mathbb{E}\Big(\textstyle\sum_{i=1}^{C} R_i \log\Big(\frac{R_i}{\hat{P}_i}\Big)\Big)$$

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

R is the variable to be optimized and can be perceived as constant wrt $E_{P(\hat{P})}(\cdot)$

$$\overset{\text{lin. of } E}{=} \min_z \ \textstyle\sum_{i=1}^{C} R_i \log R_i - R_i \, E(\log(\hat{P}_i))$$

$$\overset{\text{def } R}{=} \min_z \ \textstyle\sum_{i=1}^{C} z_i \exp(z_i) - E(\log(\hat{P}_i)) \cdot \exp(z_i)$$

and R shall be a distribution, i.e., $\forall i \ R_i \geq 0$ and $\sum_{i=1}^{C} R_i = 1$

taken care of by $R_i = \exp(z_i)$

must be constrained specifically

$$\mathcal{L}(z, \lambda) = \textstyle\sum_{i=1}^{C} z_i \exp(z_i) - E(\log(\hat{P}_i)) \exp(z_i) + \lambda\big(\sum_{i=1}^{C} \exp(z_i) - 1\big)$$

$$\frac{\partial \mathcal{L}(z, \lambda)}{\partial z_j} = (1 + z_j)\exp(z_j) - E(\log(\hat{P}_j))\exp(z_j) + \lambda \exp(z_j)$$

$$= \big(1 + z_j - E(\log(\hat{P}_j)) + \lambda\big) \underbrace{\exp(z_j)}_{>0} \overset{!}{=} 0 \qquad \Big| \cdot \exp(z_j)$$

$$\Longleftrightarrow \ 1 + z_j - E(\log(\hat{P}_j)) + \lambda = 0$$

$$\Longleftrightarrow \ z_j = E(\log(\hat{P}_j)) - 1 - \lambda$$

Plug solution into R_i:

$$R_i = \exp(z_i) = \exp\Big(E(\log(\hat{P}_i)) - 1 - \lambda\Big)$$

$$= \frac{\exp\big(E(\log(\hat{P}_i))\big)}{\exp(1 + \lambda)}$$

$$\frac{\partial \mathcal{L}(z,\lambda)}{\partial \lambda} \overset{!}{=} 0 \Rightarrow \sum_{i=1}^{c} \exp(z_i) = 1$$

taking the derivative wrt the Lagrangian multiplier λ always gives the constraint

$$\Leftrightarrow \sum_{i=1}^{c} \frac{\exp(E(\log(\hat{p}_i)))}{\exp(\lambda+1)} = 1$$

$$\Leftrightarrow \sum_{i=1}^{c} \exp(E(\log(\hat{p}_i))) = \exp(\lambda+1)$$

$$\Longrightarrow R_i = \frac{\exp(E(\log(\hat{p}_i)))}{\sum_{i=1}^{c} \exp(E(\log(\hat{p}_i)))} \qquad \square$$

(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

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

(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*.)*

Auxiliaries:

$$\mathrm{Bias}(\hat{P}) = D_{KL}(P\|R)$$
$$= \sum_i p_i \log(p_i/R_i)$$
$$= \sum_i p_i \log p_i - p_i \log R_i$$

$$\mathrm{Var}(\hat{P}) = E(D_{KL}(R\|\hat{P}))$$
$$= E\left(\sum_i R_i \log R_i - R_i \log \hat{p}_i\right)$$

$$R_i = \frac{\exp(E(\log(\hat{p}_i)))}{\exp(\lambda+1)}$$

$$\boxed{\begin{aligned} * \quad & E(\log R_i - \log \hat{p}_i) \\ &= \log R_i - E(\log(\hat{p}_i)) \\ &= E(\log(\hat{p}_i)) - (\lambda+1) - E(\log(\hat{p}_i)) \\ &= -(\lambda+1) \end{aligned}}$$

Error:

$$\mathrm{Error}(\hat{P}) = E(D_{KL}(P\|\hat{P}))$$
$$= E\left(\sum_{i=1}^{c} p_i \log(p_i/\hat{p}_i)\right)$$
$$= E\left(\sum_i p_i \log p_i - p_i \log \hat{p}_i\right)$$

$+0$
$$= E\left(\underbrace{\sum_i p_i \log p_i - p_i \log R_i}_{} + p_i \log R_i - p_i \log \hat{p}_i\right)$$

def D_{KL}
$$= E\left(D_{KL}(P\|R) + \sum_{i=1}^{c} p_i (\log R_i - \log \hat{p}_i)\right)$$

$$= D_{KL}(P\|R) + \sum_{i=1}^{c} p_i \underbrace{\left(E(\log R_i - \log \hat{p}_i)\right)}_{\overset{*}{=} -(\lambda+1)}$$

which does not depend on i

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

both are probabilities so they sum to 1

$$= \mathrm{Bias}(\hat{P}) + (-(\lambda+1)) \sum_i R_i$$

def $(-(\lambda+1))$
$$= \mathrm{Bias}(\hat{P}) + \sum_{i=1}^{c} R_i (E(\log R_i - \log \hat{p}_i))$$
$$= \mathrm{Bias}(\hat{P}) + E\left(\sum_{i=1}^{c} R_i \log R_i - R_i \log \hat{p}_i\right)$$
$$= \mathrm{Bias}(\hat{P}) + E\left(\sum_i R_i \log(R_i/\hat{p}_i)\right)$$

def $\mathrm{Var}(\hat{P})$
$$= \mathrm{Bias}(\hat{P}) + E(D_{KL}(R\|\hat{P}))$$

$$= \mathrm{Bias}(\hat{P}) + \mathbb{E}\left(\sum_i R_i \log\left(R_i / \hat{P}_i\right)\right)$$

$$\overset{\text{def. } \mathrm{Var}(\hat{P})}{=} \mathrm{Bias}(\hat{P}) + \mathbb{E}\left(D_{KL}(R \| \hat{P})\right)$$

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

$\square$