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

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

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

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

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

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

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.

$E[x_i] = \mu \qquad Var[x_i] = \sigma^2$

$$Bias(\hat{\mu}) = E[\hat{\mu} - \mu] = E[\alpha \cdot \frac{1}{N}\sum_{i=1}^{N} x_i - \frac{1}{N}\sum_{i=1}^{N} x_i] = (\alpha - 1)\mu$$

$$Var(\hat{\mu}) = Var(\alpha \cdot \frac{1}{N}\sum_{i=1}^{N} x_i) = \frac{\alpha^2}{N^2}\sum_{i=1}^{N} Var[x_i] = \frac{\alpha^2}{N^2}\cdot N \cdot \sigma^2 = \frac{\alpha^2}{N}\cdot\sigma^2$$

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

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:

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

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

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

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

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

(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.)

$$R_i = \exp(z_i) \Rightarrow \sum_{i=1}^{C} R_i = 1 = \sum_{i=1}^{C} \exp(z_i) = 1$$

$$E[D_{KL}(R||\hat{P})] = E[\sum_i R_i \cdot \log\frac{R_i}{\hat{P}_i}] = \sum_i E[R_i \cdot \log R_i] - E[R_i \cdot \log \hat{P}_i]$$

$$= \sum_i R_i \cdot \log R_i - R_i \cdot E[\log \hat{P}_i]$$

$$= \sum_i \exp(z_i)\cdot z_i - \exp(z_i)\cdot E[\log \hat{P}_i], \text{ where } \sum_{i=1}^{C} \exp(z_i) = 1$$

$$\Rightarrow \mathcal{L}(z) = \sum_i \exp(z_i)\cdot z_i - \exp(z_i)\cdot E[\log(\hat{P}_i)] + \lambda(\sum_{i=1}^{C} \exp(z_i) - 1)$$

$$= \sum_i \exp(z_i) \cdot \left[z_i - E[\log(\hat{p_i}) + \lambda \right] - \lambda$$

$$\nabla_z \mathcal{L}(z) = \sum_i \exp(z_i) \cdot [z_i - E[\log(\hat{p_i}) + \lambda + 1] = 0$$

$$\Leftrightarrow \quad z_i - E[\log(\hat{p_i})] + \lambda + 1 = 0$$

$$\Rightarrow \quad z_i = E[\log(\hat{p_i})] - \lambda - 1$$

$$\Rightarrow \quad R_i = \exp(E[\log(\hat{p_i})] - \lambda - 1)$$

$$\Rightarrow \quad R_i = \frac{\exp(E[\log(\hat{p_i})]}{\exp(1+\lambda)}$$

So we can let $\exp(1+\lambda) = \sum_j \exp(E[\hat{p_j}])$

(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

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

(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*.)*

$$E[\log R_i - \log \hat{p_i}] = E\left[\log \frac{\exp(E[\log(\hat{p_i})]}{\exp(1+\lambda)} - \log \hat{p_i}\right]$$

$$= E[E[\log \hat{p_i}] - (1+\lambda) - \log \hat{p_i}]$$

$$= -(1+\lambda)$$

$$\text{Bias}(\hat{p}) + \text{Var}(\hat{p}) = D_{KL}(p||R) + E[D_{KL}(R||\hat{p})]$$

$$= \sum_i p_i \cdot \ln \frac{p_i}{R_i} + E\left[\sum_i R_i \cdot \ln \frac{R_i}{\hat{p}_i}\right]$$

$$= \sum_i p_i \cdot \ln \frac{p_i}{R_i} + E\left(\sum_i R_i (\log R_i - \log \hat{p_i})\right)$$

$$= \sum_i p_i \cdot \ln \frac{p_i}{R_i} + E\left(\sum_i p_i (\log R_i - \log \hat{p_i})\right)$$

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

$$= E\left(\sum_i p_i \log p_i - p_i \log \hat{p}_i \right)$$

$$= E\left(DL(p \,\|\, \hat{p}) \right)$$

$$= \text{error}\,(\hat{p})$$