Machine Learning Sheet 9

1. **Bias and Variance of Mean Estimators**

a)

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

$$\begin{aligned}
\mathbf{Bias}(\hat{\mu}) &= \mathbb{E}[\hat{\mu}] - \mu \\
&= \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N} X_i] - \mu \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[X_i] - \mu \\
&= \frac{1}{N} \cdot N\mu - \mu \\
&= 0 \\
\mathbf{Var}(\hat{\mu}) &= \mathbb{E}[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2] \\
&= \mathbb{E}[(\hat{\mu} - \mu)^2] \\
&= \mathbb{E}[(\frac{1}{N}\sum_{i=1}^{N} X_i - \mu)^2] \\
&= \frac{1}{N^2}\mathbb{E}[\sum_{i=1}^{N}\sum_{j=1}^{N}(X_i - \mu)(X_j - \mu)] \\
&= \frac{1}{N^2}\left(\sum_{i=1}^{N}\mathbf{Var}(X_i) + \sum_{i \neq j}\mathbf{Cov}(X_i, X_j)\right)
\end{aligned}$$

Notice that, for $i \neq j$, X_i and X_j are independent and uncorrelated, which means $\mathbf{Cov}(X_i, X_j) = 0$

$$\begin{aligned}
\mathbf{Var}(\hat{\mu}) &= \frac{1}{N^2}\sum_{i=1}^{N}\mathbf{Var}(X_i) \\
&= \frac{1}{N^2} \cdot N\sigma^2 \\
&= \frac{\sigma^2}{N}
\end{aligned}$$

$$\mathbf{Error}(\hat{\mu}) = \mathbf{Bias}(\hat{\mu})^2 + \mathbf{Var}(\hat{\mu}) = \frac{\sigma^2}{N}$$

b)

$$\begin{aligned}
\hat{\mu} &= 0 \\
\mathbf{Bias}(\hat{\mu}) &= \mathbb{E}[\hat{\mu} - \mu] = -\mu \\
\mathbf{Var}(\hat{\mu}) &= \mathbb{E}[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2] \\
&= \mathbb{E}[(0 - 0)^2] \\
&= 0
\end{aligned}$$

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

2. **Bias-Variance Decimposition for Regression**

a)

$$
\begin{aligned}
\mathbf{Bias}(\hat{f}(x)) &= \mathbb{E}[\hat{f}(x) - f(x)] \\
&= \mathbb{E}[\hat{f}(x)] - f(x) \\
\mathbf{Var}(\hat{f}(x)) &= \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2] \\
&= \mathbb{E}[\hat{f}^2(x)] - 2\mathbb{E}[\hat{f}(x)] \cdot \mathbb{E}[\hat{f}(x)] + \mathbb{E}[\hat{f}(x)]^2 \\
&= \mathbb{E}[\hat{f}^2(x)] - \mathbb{E}[\hat{f}(x)]^2 \\
\mathbf{Error}(\hat{f}(x)) &= \mathbb{E}[(\hat{f}(x) - f(x))^2] \\
&= \mathbb{E}[\hat{f}^2(x) - 2\hat{f}(x)f(x) + f^2(x)] \\
&= \mathbb{E}[\hat{f}^2(x)] - 2\mathbb{E}[\hat{f}(x)]f(x) + f^2(x) \\
&= \Big(\mathbb{E}[\hat{f}^2(x)] - \mathbb{E}[\hat{f}(x)]^2\Big) + \Big(\mathbb{E}[\hat{f}(x)]^2 - 2\mathbb{E}[\hat{f}(x)]f(x) + f^2(x)\Big) \\
&= \mathbf{Var}(\hat{f}(x)) + \mathbf{Bias}(\hat{f}(x))^2
\end{aligned}
$$

3. **Bias-Variace Decomposition for Classification**

a)Use Lagrange Multiplier:

$$
\begin{aligned}
\mathcal{L} &= \mathbb{E}[\sum_{i=1}^{C} R_i log \frac{R_i}{\hat{P}_i}] - \lambda(\sum_{i=1}^{C} R_i - 1) \\
\frac{\partial \mathcal{L}}{\partial R_i} &= 1 - logR_i + \mathbb{E}[log\hat{P}_i] - \lambda = 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= \sum_{i=1}^{C} R_i - 1 = 0 \\
\Rightarrow \qquad R_i &= exp(1 - \lambda + \mathbb{E}[log\hat{P}_i]) \\
\sum_{i=1}^{C} exp((1 - \lambda &+ \mathbb{E}[log\hat{P}_i])) = 1 \\
\Rightarrow \qquad R_i &= \frac{exp(\mathbb{E}[log\hat{P}_i])}{\sum_{j=1}^{C} exp(\mathbb{E}[log\hat{P}_j])}
\end{aligned}
$$

b)

$$\mathbf{Error}(\hat{P}) = \mathbb{E}[D_{KL}(P||\hat{P})] = \mathbb{E}[\sum_{i=1}^{C} P_i log \frac{P_i}{\hat{P}_i}] = \sum_{i=1}^{C} P_i log P_i - \mathbb{E}[\sum_{i=1}^{C} P_i log \hat{P}_i] = \sum_{i=1}^{C} P_i log P_i - \sum_{i=1}^{C} P_i \mathbb{E}[log\hat{P}_i]$$

$$\mathbf{Bias}(\hat{P}) = D_{KL}(P||R) = \sum_{i=1}^{C} P_i log P_i - \sum_{i=1}^{C} P_i log R_i$$

$$\mathbf{Var}(\hat{P}) = \mathbb{E}[D_{KL}(R||\hat{P})] = \mathbb{E}[\sum_{i=1}^{C}(R_i log R_i - R_i log \hat{P}_i)]$$

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

$$
\begin{aligned}
\mathbf{Bias}(\hat{P})+\mathbf{Var}(\hat{P}) &= \sum_{i=1}^{C} P_i log P_i - \sum_{i=1}^{C} P_i log R_i + \mathbb{E}[\sum_{i=1}^{C}(R_i log R_i - R_i log \hat{P}_i)] \\
&= \sum_{i=1}^{C} P_i log P_i - \mathbb{E}[\sum_{i=1}^{C}(P_i log R_i - R_i log R_i + R_i log \hat{P}_i)] \\
&= \sum_{i=1}^{C} P_i log P_i - \sum_{i=1}^{C} \mathbb{E}[(P_i log R_i - R_i log R_i + R_i log \hat{P}_i)] \\
&= \sum_{i=1}^{C} P_i log P_i - \sum_{i=1}^{C} \left(P_i \mathbb{E}[log \frac{exp(\mathbb{E}[log \hat{P}_i])}{\sum_{j=1}^{C} exp(\mathbb{E}[log \hat{P}_i])}] - \mathbb{E}[R_i log R_i - R_i log \hat{P}_i)] \right) \\
&= \sum_{i=1}^{C} P_i log P_i - \sum_{i=1}^{C} \left(P_i \mathbb{E}[log \hat{P}_i] - P_i \mathbb{E}[log \sum_{j=1}^{C} exp(\mathbb{E}[log \hat{P}_i])] - \mathbb{E}[R_i log R_i - R_i log \hat{P}_i)] \right) \\
&= \sum_{i=1}^{C} P_i log P_i - \sum_{i=1}^{C} P_i \mathbb{E}[log \hat{P}_i] + \sum_{i=1}^{C} \left(P_i \mathbb{E}[log \sum_{j=1}^{C} exp(\mathbb{E}[log \hat{P}_i])] + \mathbb{E}[R_i log R_i - R_i log \hat{P}_i)] \right)
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{i=1}^{C} \left(P_i \mathbb{E}[log \sum_{j=1}^{C} exp(\mathbb{E}[log \hat{P}_i])] + \mathbb{E}[R_i log R_i - R_i log \hat{P}_i)] \right) \\
=&\sum_{i=1}^{C} \left(P_i \mathbb{E}[log \sum_{j=1}^{C} exp(\mathbb{E}[log \hat{P}_i])] + \mathbb{E}[R_i \mathbb{E}[log \hat{P}_i]] - \mathbb{E}[R_i log \sum_{j=1}^{C} exp \mathbb{E}[log \hat{P}_i] - R_i log \hat{P}_i) \right) \\
=&\sum_{i=1}^{C} \left(P_i \mathbb{E}[log \sum_{j=1}^{C} exp(\mathbb{E}[log \hat{P}_i])] - \mathbb{E}[R_i log \sum_{j=1}^{C} exp \mathbb{E}[log \hat{P}_i] \right) \\
=&\sum_{i=1}^{C} \mathbb{E} \left[P_i log \sum_{j=1}^{C} exp(\mathbb{E}[log \hat{P}_i]) - R_i log \sum_{j=1}^{C} exp \mathbb{E}[log \hat{P}_i] \right] \\
=&0
\end{aligned}
$$

$$
\mathbf{Bias}(\hat{P})+\mathbf{Var}(\hat{P}) = \sum_{i=1}^{C} P_i log P_i - \sum_{i=1}^{C} P_i \mathbb{E}[log \hat{P}_i] = \mathbf{Error}(\hat{P})
$$