COMS 4771 Machine Learning (2018 Fall)
Homework 1

Jing Qian - jq2282@columbia.edu

October 5, 2018

Problem 2

(iii).

$$\begin{aligned}
\mathbb{E}[x] &= \frac{1}{n}\sum_{i=1}^{n}[1 \times b + 0 \times (1-b)] \\
&= \frac{1}{n}\sum_{i=1}^{n} b \\
&= b, \\
\mathbb{E}[x^2] &= \frac{1}{n}\sum_{i=1}^{n}[1^2 \times b + 0^2 \times (1-b)] \\
&= \frac{1}{n}\sum_{i=1}^{n} b \\
&= b, \\
\mathrm{Var}[x] &= \mathbb{E}[x^2] - \mathbb{E}[x]^2 \\
&= b - b^2.
\end{aligned} \tag{1}$$

(iv). Using the invariance property of MLE shown in Problem 1(ii), the MLE for the coin's variance is the variance function of the MLE bias $\hat{b}$. From subproblem (i), we get the MLE bias $\hat{b} = \frac{\sum_{i=1}^{n} x_i}{n}$. From subproblem (ii), we get the variance of this coin is $\mathrm{Var}[x] = b - b^2$. So the MLE for the coin's variance is :

$$\hat{\mathrm{Var}}[x] = \hat{b} - \hat{b}^2 = \frac{\sum_{i=1}^{n} x_i}{n} - (\frac{\sum_{i=1}^{n} x_i}{n})^2. \tag{2}$$

(v).

(vi). When the parameter b has uniform distribution, MAP estimate equals MLE.
From the definition, the MLE and MAP estimations of parameter b are:

$$\begin{aligned} b_{\text{MLE}} &= \arg\ \max_b \prod_{i=1}^{N} P(\overrightarrow{x_i}|b) \\ &= \arg\ \max_b \sum_{i=1}^{N} \log P(\overrightarrow{x_i}|b), \\ b_{\text{MAP}} &= \arg\ \max_b \prod_{i=1}^{N} P(b|\overrightarrow{x_i}) \\ &= \arg\ \max_b \prod_{i=1}^{N} P(\overrightarrow{x_i}|b)P(b) \\ &= \arg\ \max_b (\log P(b) + \sum_{i=1}^{N} \log P(\overrightarrow{x_i}|b)). \end{aligned} \tag{3}$$

The difference between two estimations is the $\log P(b)$ term in b_{MAP}. To make $b_{\text{MLE}} = b_{\text{MAP}}$, $\arg\ \max_b \log P(b)$ must be zero, which means $P(b)$ is constant and hence is uniform distribution. In other words, MAP estimate equlas MLE when $P(b)$ is a uniform distribution.

Problem 3.2

Let f be the median of y given x. Then f would be the optimal predictor if we have $Q(g) - Q(f) \geq 0$ for any x in the domain.

If $f < g$:

$$\begin{aligned} \mathbb{E}[|g-y|] - \mathbb{E}[|f-y|] &= \mathbb{E}[|g-y| - |f-y|] \\ &= \Pr[y \leq f](|g-y| - |f-y|) + \Pr[y > f](|g-y| - |f-y|) \\ &= \Pr[y \leq f](g - y - f + y) + \Pr[y > f](|y-g| - |y-f|) \\ &\geq \Pr[y \leq f](g-f) + \Pr[y > f][-(g-f)] \\ &= (g-f)[\Pr[y \leq f] - \Pr[y > f]] \\ &\geq 0 \end{aligned} \tag{4}$$

according to the property of median. On the other hand, if $f > g$, similarly, we have:

$$\begin{aligned} \mathbb{E}[|g-y|] - \mathbb{E}[|f-y|] &= \mathbb{E}[|g-y| - |f-y|] \\ &= \Pr[y \geq f](|g-y| - |f-y|) + \Pr[y < f](|g-y| - |f-y|) \\ &\geq \Pr[y \geq f](f-g) + \Pr[y < f][-(f-g)] \\ &= (f-g)[\Pr[y \geq f] - \Pr[y < f]] \\ &\geq 0 \end{aligned} \tag{5}$$

Since $\mathbb{E}[|g-y|] - \mathbb{E}[|f-y|] \geq 0$ at any given x, $Q(g) \geq Q(f)$, the median of y is the optimal predictor.

Problem 5

For every training set with $(x_1, y_1), \cdots, (x_n, y_n)$ i.i.d.samples, we could find one unique w to minimize its training error $\mathcal{R}$. That is to say,

$$\hat{w} = \text{arg min}\hat{\mathcal{R}}(w) = \text{arg min } \frac{1}{n}\sum_{i=1}^{n}(w \cdot x_i - y_i)^2. \tag{6}$$

For any w, we have $\hat{\mathcal{R}}(\hat{w}) \leq \hat{\mathcal{R}}(w)$.

Similarly, for another i.i.d. random sample consisted of $(\tilde{x_1}, \tilde{y_1}), \cdots, (\tilde{x_n}, \tilde{y_n})$:

$$\tilde{w} = \text{arg min}\tilde{\mathcal{R}}(w) = \text{arg min } \frac{1}{n}\sum_{i=1}^{n}(w \cdot \tilde{x}_i - \tilde{y}_i)^2. \tag{7}$$

For any w, we have $\tilde{\mathcal{R}}(\tilde{w}) \leq \tilde{R}(w)$.Here the inequality holds for $\hat{w}$ since it holds for any w, $\tilde{\mathcal{R}}(\tilde{w}) \leq \tilde{\mathcal{R}}(\hat{w})$.

Since both sets are i.i.d. samples from the same domain:

$$\begin{aligned} \mathbb{E}[\hat{\mathcal{R}}(\hat{w})] &= \min\ \mathbb{E}[(w \cdot x - y)^2] \\ \mathbb{E}[\tilde{\mathcal{R}}(\tilde{w})] &= \min\ \mathbb{E}[(w \cdot x - y)^2] \end{aligned} \tag{8}$$

So the expectations of training error of two i.i.d. random samples equal: $\mathbb{E}[\hat{\mathcal{R}}(\hat{w})] = \mathbb{E}[\tilde{\mathcal{R}}(\tilde{w})]$. Then we have:

$$\mathbb{E}[\hat{\mathcal{R}}(\hat{w})] = \mathbb{E}[\tilde{\mathcal{R}}(\tilde{w})] \leq \mathbb{E}[\tilde{\mathcal{R}}(\hat{w})]. \tag{9}$$

Since the set $(\tilde{x_1}, \tilde{y_1}), \cdots, (\tilde{x_n}, \tilde{y_n})$ is a i.i.d.random sample from the domain, the inequality above holds for the generalization of any i.i.d.random samples with squared error $\mathcal{R}$. In other words:

$$\mathbb{E}[\hat{\mathcal{R}}(\hat{w})] \leq \mathbb{E}[\mathcal{R}(\hat{w})]. \tag{10}$$