CSCI 6100 Machine Learning From Data
Fall 2018

HOMEWORK 8

Daniel Southwick
661542908
southd@rpi.edu

Exercise 4.3

a) If $\mathcal{H}$ is fixed and we increase the complexity of f, then the deterministic noise will go up in general. There is a higher tendency to overfit.

b) If f is fixed and we decrease the complexity of $\mathcal{H}$, then the deterministic noise will go up in general. There is a lower tendency to overfit.

Exercise 4.5

a) When $\Gamma = \mathrm{I}$, where I is an identity matrix with dimension $Q + 1$, then we can obtain a constraint in the form of $\sum_{q=0}^{Q} w_q^2 \leq C$:

$$w^T\Gamma^T\Gamma w = w^T\mathrm{I}^T\mathrm{I}w = w^T w = \sum_{q=0}^{Q} w_q^2$$

b) When $\Gamma = [1, 1, ..., 1]$, we can obtain a constraint in the form of $(\sum_{q=0}^{Q} w_q)^2 \leq C$:

$$w^T\Gamma^T\Gamma w = [w_0 \dots w_q] \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} [1 \dots 1] \begin{bmatrix} w_0 \\ \vdots \\ w_q \end{bmatrix} = \sum_{q=0}^{Q} w_q \sum_{q=0}^{Q} w_q = \left(\sum_{q=0}^{Q} w_q \right)^2$$

Exercise 4.6

The hard-order constraint should be more useful in terms of binary classification. Let w^* denote the optimal weight vector without constraints, and if the constraint $w^T w \leq C$ was added, the new optimal vector became $w = w^* * \frac{C}{w^{*T}w^*}$. So $sign(w^{*T}x) = sign(w^* * \frac{C}{w^{*T}w^*}x)$. So the constraint gets satisfied and for binary classification, we only want to know the sign of $w^T x$, the larger weight or smaller weight cannot affect the penalty.

Exercise 4.7

a) Since (x_n, y_n) is independent from data (x_m, y_m):

$$\begin{aligned}
\sigma^2_{\text{val}} &= \text{Var}_{\mathcal{D}_{\text{val}}}\left[E_{\text{val}}(g^-)\right] \\
&= \text{Var}_{\mathcal{D}_{\text{val}}}\left[\frac{1}{K}\sum_{x_n\in\mathcal{D}_{\text{val}}} e(g^-(x_n), y_n)\right] \\
&= \frac{1}{K^2}\left[\sum_{x_n\in\mathcal{D}_{\text{val}}} Var(e(g^-(x_n), y_n)) + \sum_{x_n, x_m\in\mathcal{D}_{\text{val}, n\neq m}} COV(e(g^-(x_n), y_n), e(g^-(x_m), y_m))\right] \\
&= \frac{1}{K^2}\left[\sum_{x_n\in\mathcal{D}_{\text{val}}} Var(e(g^-(x_n), y_n))\right] \\
&= \frac{1}{K^2}\cdot K\cdot \text{Var}\left[e(g^-(x), y)\right] \\
&= \frac{1}{K^2}\cdot K\cdot \sigma^2(g^-) \\
&= \frac{1}{K}\sigma^2(g^-)
\end{aligned}$$

b) Since $e(g^-(x), y) = [g^-(x) \neq y]$

$$\begin{aligned}
\mathbb{P}\left[e(g^-(x), y) = 1\right] &= p \\
\mathbb{P}\left[e(g^-(x), y) = 0\right] &= 1 - p \\
E(e(g^-(x), y)) &= 1 * p + 0 * (1-p) = p
\end{aligned}$$

$$\begin{aligned}
\text{Var}_{\mathcal{D}_{\text{val}}}\left[E_{\text{val}}(g^-)\right] &= \text{Var}_{\mathcal{D}_{\text{val}}}\left[\frac{1}{K}\sum_{x_n\in\mathcal{D}_{\text{val}}} e(g^-(x_n), y_n)\right] \\
&= \frac{1}{K^2}\cdot \text{Var}\left[e(g^-(x), y)\right] \\
&= \frac{1}{K}\left[p(1-p)^2 + (1-p)(0-p)^2\right] \\
&= \frac{p-p^2}{K}
\end{aligned}$$

$$\mathbb{P}\left[g^-(\mathbf{x}) \neq y\right] = \frac{1}{K}\mathbb{P}\left[g^-(\mathbf{x}) \neq y\right]^2$$

c) We can modify the result from part b) to:

$$\begin{aligned}
\text{Var}_{\mathcal{D}_{\text{val}}}\left[E_{\text{val}}(g^-)\right] &= \frac{p-p^2}{K} \\
&= \frac{1}{K}\left[-(p-\frac{1}{2})^2 + \frac{1}{4}\right]
\end{aligned}$$

So the maximum of σ_{val}^2 is $\dfrac{1}{4K}$, thus $\sigma_{val}^2 \leq \dfrac{1}{4K}$

d) $e(g^-(x), y) = (g^-(x) - y)^2$

$$\begin{aligned}
\text{Var}_{\mathcal{D}_{\text{val}}}\left[E_{\text{val}}(g^-)\right] &= \text{Var}_{\mathcal{D}_{\text{val}}}\left[\frac{1}{K}\sum_{x_n \in \mathcal{D}_{\text{val}}} e(g^-(x_n), y_n)\right] \\
&= \frac{1}{K^2} \cdot \text{Var}\left[e(g^-(x), y)\right] \\
&= \frac{1}{K}\left[p(1-p)^2(g^-(x) - y)^4 + (1-p)(0 - p(g^-(x) - y)^2)^2\right] \\
&= \frac{p - p^2}{K}(g^-(x) - y)^4
\end{aligned}$$

Since the square error is unbounded, thus there is no uniform upper bound for $Var\left[E_{val}(g^-)\right]$

e) The $\sigma^2(g^-)$ expected to be higher. Since fewer data points indicates more possible to overfit, then theoretically the out of sample error should be larger and the validation error is expected to increase.

f) Since $\sigma_{\text{val}}^2 = \dfrac{1}{K}\sigma^2(g^-)$(from part a)), increasing the size of validation set means increasing K. but it also results in decreasing in size of training set, which leads to worse g that increases $\sigma^2(g^-)$. Thus if we estimate $E_{out}(g)$ with $E_{out}(g^-)$, then increasing the size of the validation set can result in a better or worse estimate of E_{out}, either case could happen.

Exercise 4.8

Yes, E_m is an unbiased estimate for the out of sample error $E_{out}(g_m^-)$, since the selection of g_m^- depends only on training data set, which does not include validation data points.

Problem 4.26

a)

$$Z = \begin{bmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_N^T \end{bmatrix}, \; Z^T = [z_1 z_2 \ldots z_N]$$

$$Z^TZ = [z_1 z_2 \dots z_N] \begin{bmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_N^T \end{bmatrix} = \sum_{n=1}^{N} z_n z_n^T$$

$$Z^Ty = [z_1 z_2 \dots z_N]\, y = \sum_{n=1}^{N} z_n y_n$$

$$H(\lambda) = Z(Z^TZ + \lambda\Gamma^T\Gamma)^{-1}Z^T = \begin{bmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_N^T \end{bmatrix} (Z^TZ + \lambda\Gamma^T\Gamma)^{-1}\, [z_1 z_2 \dots z_N]$$

$$H_{nm}(\lambda) = z_n^T(Z^TZ + \lambda\Gamma^T\Gamma)^{-1}z_m = z_n^T A^{-1} z_m$$

And when (z_n, y_n) is left out, from the previous terms $Z^TZ \to Z^TZ - z_n z_n^T$ and $Z^Ty \to Z^TY - z_n y_n$

b)

Since $(A - xx^T)^{-1} = A^{-1} + \dfrac{A^-1xx^TA^{-1}}{1 - x^TA^{-1}x}$, we replace x by z_n.

$$\begin{aligned} w &= (Z^TZ + \lambda\Gamma^T\Gamma)^{-1}Z^Ty = A^{-1}Z^Ty \\ w_n^- &= (Z^TZ + \lambda\Gamma^T\Gamma - z_n z_n^T)^{-1}(Z^Ty - z_n y_n) \\ &= (A - z_n z_n^T)^{-1}(Z^Ty - z_n y_n) \\ &= (A^{-1} + \frac{A^{-1}z_n z_n^T A^{-1}}{1 - z_n^T A^{-1} z_n})(Z^Ty - z_n y_n) \end{aligned}$$

c)

$$\begin{aligned} w_n^- &= (A^{-1} + \frac{A^{-1}z_n z_n^T A^{-1}}{1 - z_n^T A^{-1} z_n})(Z^Ty - z_n y_n) \\ &= A^{-1}Z^T + \frac{A^{-1}z_n z_n^T A^{-1}}{1 - z_n^T A^{-1} z_n}Z^Ty - A^{-1}z_n y_n - \frac{A^{-1}z_n z_n^T A^{-1}}{1 - z_n^T A^{-1} z_n} z_n y_n \\ &= w + \frac{A^{-1}z_n z_n^T w}{1 - H_{nn}} - A^{-1}z_n y_n - \frac{A^{-1}z_n H_{nn} y_n}{1 - H_{nn}} \\ &= w + \frac{A^{-1}z_n \hat{y_n}}{1 - H_{nn}} - \frac{A^{-1}z_n y_n}{1 - H_{nn}} \\ &= w + \frac{\hat{y_n} - y_n}{1 - H_{nn}} A^{-1} z_n \end{aligned}$$

d)

$$\begin{aligned}
w_n^- &= w + \frac{y_n - \hat{y_n}}{1 - H_{nn}} A^{-1} z_n \\
z_n^T w_n^- &= z_n^T (w + \frac{y_n - \hat{y_n}}{1 - H_{nn}} A^{-1} z_n) \\
&= \hat{y_n} + \frac{\hat{y_n} - y_n}{1 - H_{nn}} z_n^T A^{-1} z_n \\
&= \hat{y_n} + \frac{\hat{y_n} - y_n}{1 - H_{nn}} H_{nn} \\
&= \frac{\hat{y_n} - H_{nn} y_n}{1 - H_{nn}}
\end{aligned}$$

e)

$$\begin{aligned}
e_n &= (z_n^T w_n^- - y_n)^2 \\
&= (\frac{\hat{y_n} - H_{nn} y_n}{1 - H_{nn}} - y_n)^2 \\
&= (\frac{\hat{y_n} - y_n}{1 - H_{nn}})^2 \\
E_{cv} &= \frac{1}{N} \sum_{n=1}^{N} e_n \\
&= \frac{1}{N} \sum_{n=1}^{N} (\frac{\hat{y_n} - y_n}{1 - H_{nn}})^2
\end{aligned}$$