Homework 8
CSCI4100

Han Hai
Rin:661534083
haih2@rpi.edu

October 28th 2018

1. **[200 points] Exercise 4.3** Deterministic noise depends on H, as some models approximate f better than others.

   (a) Assume H is fixed and we increase the complexity of f. Will deterministic noise in general go up or down? Is there a higher or lower tendency to overfit?
   Solution: If H is fixed and we increase the complexity of f, deterministic noise will go up, because there are more areas that we cannot model. And there is a lower tendency to overfit, because the target function becomes a more complex model, the original H will be more likely to underfit and less likely to overfit.

   (b) Assume f is fixed and we dcrease the complexity of H. Will deterministic noise in general go up or down? Is there a higher or lower tendency to overfit?
   Solution: deterministic noise will go up since H becomes simpler and it cannot model the complexity in f, and there is less tendency to overfit, since hypothesis model becomes simpler

2. **[200 points] Exercise 4.5 [Tikhonov regularizer]**
   A more gneral soft constraint is the Tikhonov regularization constraint:
   $w^T\Gamma^T\Gamma w \leq C$

   (a) What should Γ be to obtain the constraint $\sum_{q=0}^{Q} w_q^2 \leq C$
   Solution:
   When $\Gamma = I$
   $w^T\Gamma^T\Gamma w = w^T w = \sum_{q=0}^{Q} w_q^2 \leq C$

   (b) What should Γ be to obtain the constraint $(\sum_{q=0}^{Q} w_q)^2 \leq C$
   Solution:
   Let Γ be a $(d+1) \times (d+1)$ matrix with first row 1s and the rest 0s.
   Γw becomes $(\sum_{q=0}^{Q} w_q), 0, 0, 0...)^T$
   So, $w^T\Gamma^T\Gamma w = (\sum_{q=0}^{Q} w_q)^2 \leq C$

3. **[100points] Exercise 4.6**
   We have seen both the hard-order constraint and the soft-order constraint. Which do you expect to be more useful for binary classification using the perceptron model?
   Hint: $sign(w^T x) = sign(\alpha w^T x)$
   Solution:
   I think hard-order constraint is more useful in this context. A soft-order constraint in this case is similar to what the hint describes, for $\alpha > 0$ weight won't affect the classification much. In contrary, setting weights to zero directly forbid the possibility of overfitting in this weight. Therefore, I believe

hard-order constraint is better.

4. **[100 points] Exercise 4.7** Fix g^-(learned from D_{train}) and define $\sigma_{val}^2 = Var_{D_{val}}[E_{val}(g^-)]$ We consider how σ_{val}^2 depends on K. Let $\sigma^2(g^-) = Var_x[e(g^-(x), y)]$ be the pointwise variance in the out-of-sample error of g^-.

   (a) Show that $\sigma_{val}^2 = \frac{1}{K}\sigma^2(g^-)$
   $\sigma_{val}^2 = Var_{D_{val}}[E_{val}(g^-)]$

   $=\frac{1}{K}Var_{D_{val}}[e(g^-(x), y)]$

   $=\frac{1}{K}\sigma^2(g^-)$

   (b) In a classification problem, where $e(g-(x), y) = [g-(x) \neq y]$ express σ_{val}^2 in terms of $P[g^-(x) \neq y]$
   $\sigma_{val}^2 = \frac{1}{K}\sigma^2(g^-)$
   $=\frac{1}{K}Var_{D_{val}}[E_{val}(g^-)]$
   $=\frac{1}{K}(E_x[e(g-(x), y) - (E_x[e(g-(x), y)]])^2$
   $=\frac{1}{K}(Ex(1 - P[g-(x) \neq y] - P[g-(x) \neq y]^2)$
   $=\frac{1}{K}(1 - P[g^-(x \neq y)])(P[g^-(x \neq y))$

   (c) Show that for any g- in a classification problem,$\sigma_{val}^2 \leq \frac{1}{4K}$
   for any $x \leq 1$ $(1-x)x \leq \frac{1}{4}$
   therefore $\sigma_{val}^2 = \frac{1}{K}(1-P)P \leq \frac{1}{4K}$

   (d) Is there a uniform upper bound for variance similar to c in the case of regression with squred error $e(g(x), y) = (g(x) - y)^2$
   No, since squared error is unbounded, $Var[E_{val}(g^-)] = e^2 - e^2$ is also unbounded

   (e) For regression with squred error, if we train using fewer points to get g^- do we expect $\sigma^2(g-)$ to be higher or lower?
   I expect it to be higher, a smaller training set will result in a higher expected error in terms of mean, and higher mean result in higher variance

   (f) Conclude that increasing the size of the validation set can result in a better or a worse estimate of E_{out}
   It depends, a higher validation set size reduce σ_{val}^2, and it increases $\sigma_{g^-}^2$. There is a trade off, so finding the balance is key.

5. **[Exercise 4.8]**
   Is E_m an unbiased estimate for the out of sample error $E_{out}(g_m^-)$?
   Yes, $E_m = E_{val}(g_m^-)$ and validation set is independent of training. Therefore it must be unbiased.