

ASSIGNMENT 4, Solutions

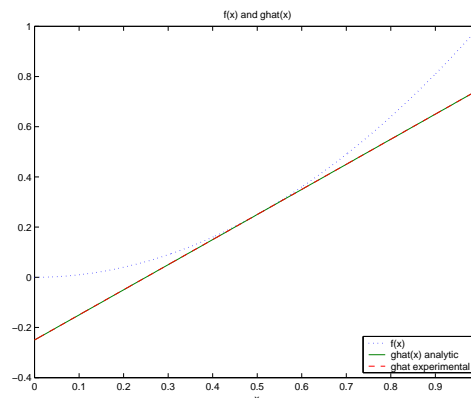
1 (300pts) Computation of BIAS and VARIANCE

- (a) The idea is to compute the expectations using averages. Generate many data sets. For each data set, compute the learned function. The average of these functions gives $\hat{g}(x)$. The average integrated squared error (which can be computed numerically) is an estimate of R and the BIAS can be numerically computed once $\hat{g}(x)$ is computed by computing the integrated squared difference between \hat{g} and f . The simplest way to compute VAR, is to rerun the experiment or to keep all the learned g 's in order to compute the average of the integrated squared difference between each g and \hat{g} . This requires extra running time or humongous memory as to get accurate estimates, one needs to run about 10^6 data sets. A tricky approach is to realize that

$$\begin{aligned} E_{D_N}[(g(x, D_N) - \hat{g}(x))^2] &= E_{D_N}[g(x, D_N)^2] - 2\hat{g}(x)E_{D_N}[g(x, D_N)] + \hat{g}(x)^2 \\ &= E_{D_N}[g(x, D_N)^2] - \hat{g}(x)^2 \end{aligned}$$

thus one could also compute the average squared function \hat{g}^2 just as one computes the average function \hat{g} , and then use that to compute the VAR by taking the relevant numerical integrals.

- (b) Using 10^6 data sets, numerically I obtained $R = 0.0333567$, $BIAS = 0.0125279$, $VAR = 0.0208288$. The plot of $f(x)$ and $\hat{g}(x)$ are shown below.



- (c) (extra credit, 50pts) It is not hard to show that the line that fits the 2 points $\{(x_1, x_1^2), (x_2, x_2^2)\}$ is given by $g(x) = (x_1 + x_2)x - x_1x_2$. Taking the expectation with respect to x_1, x_2 gives $\hat{g}(x) = x - \frac{1}{4}$. This function is also plotted in the above figure, and is virtually identical to the empirically obtained $\hat{g}(x)$. R , BIAS and VAR are then given by integrals as follows.

$$\begin{aligned} R &= \int_0^1 dx \int_0^1 dx_1 \int_0^1 dx_2 ((x_1 + x_2)x - x_1x_2 - x^2)^2 = \frac{1}{30} = 0.033333333 \dots \\ BIAS &= \int_0^1 dx (x^2 - x + \frac{1}{4})^2 = \frac{1}{80} = 0.0125. \\ VAR &= \int_0^1 dx \int_0^1 dx_1 \int_0^1 dx_2 ((x_1 + x_2)x - x_1x_2 - x + \frac{1}{4})^2 = \frac{1}{48} = 0.0208333333 \dots \end{aligned}$$

These exact values can be compared to the experimentally obtained ones in (b).

2 (350pts) Weight Decay Trades Bias for Variance

(a) Suppose we estimate m by w . We want to maximize $P[m = w|D_N]$.

$$P[m = w|D_N] = \frac{P[D_N|m = w]P[m = w]}{P[D_N]}$$

and since $P[m = w]$ is uniform and $P[D_N]$ is independent of w , we can pick w so that it maximizes $P[D_N|m = w] \propto \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - w)^2\right]$. Equivalently we minimize over w

$$\mathcal{E}_0 = \frac{1}{N} \sum_{i=1}^N (y_i - w)^2$$

Setting $\frac{\partial \mathcal{E}_0}{\partial w} = 0$, we see that $\frac{1}{N} \sum_{i=1}^N (w - y_i) = 0 \Rightarrow \hat{m}_0 = \frac{1}{N} \sum_{i=0}^N y_i = \frac{1}{N} \sum_{i=0}^N (m + \epsilon_i)$

(b) Let $\mathcal{E} = \mathcal{E}_0 + \lambda w^2$. Computing $\frac{\partial \mathcal{E}}{\partial w} = \frac{\partial \mathcal{E}_0}{\partial w} + 2\lambda w = 2(w - \hat{m}_0) + 2\lambda w$ and setting this to zero immediately yields

$$\hat{m} = \frac{\hat{m}_0}{1 + \lambda}$$

(c)

$$R(\lambda) = E_\epsilon \left[\left(\frac{\hat{m}_0}{1 + \lambda} - m \right)^2 \right] = \frac{E_\epsilon [\hat{m}_0^2]}{(1 + \lambda)^2} - \frac{2m E_\epsilon [\hat{m}_0]}{1 + \lambda} + m^2 \quad (1)$$

Thus we need $E_\epsilon [\hat{m}_0^2]$ and $E_\epsilon [\hat{m}_0]$. Since $E[\epsilon_i] = 0$ it is easy to see that $E_\epsilon [\hat{m}_0] = m$

$$\begin{aligned} E_\epsilon [\hat{m}_0^2] &= \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (m + \epsilon_i)(m + \epsilon_j) = m^2 + \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N E_\epsilon [\epsilon_i \epsilon_j] + \frac{2m}{N^2} \sum_{i=1}^N E_\epsilon [\epsilon_i] \\ &= m^2 + \frac{\sigma^2}{N^2} \sum_{i=1}^N \sum_{j=1}^N \delta_{ij} = m^2 + \frac{\sigma^2}{N} \end{aligned}$$

Now, plugging into (1) gives $R(\lambda) = \frac{\lambda^2 m^2}{(1 + \lambda)^2} + \frac{\sigma^2}{N(1 + \lambda)^2}$

(d) Setting $\frac{d}{d\lambda} R(\lambda) = 0$ and solving for λ_{opt} one finds that $\lambda_{opt} = \sigma^2/Nm^2$. Noting that $R(0) = \sigma^2/N$ we see that $m^2 = R(0)/\lambda_{opt}$. Hence, $R(\lambda_{opt}) = \frac{1}{(1 + \lambda_{opt})^2} \left(\lambda_{opt}^2 \frac{R(0)}{\lambda_{opt}} + R(0) \right)$, which yields

$$R(\lambda_{opt}) = \frac{R(0)}{1 + \lambda_{opt}} < R(0) \text{ for } \lambda_{opt} > 0$$

(e) $E_\epsilon [m - \hat{m}] = m - \frac{E_\epsilon [\hat{m}_0]}{1 + \lambda_{opt}} = m - \frac{m}{1 + \lambda_{opt}} \Rightarrow bias = \frac{\lambda_{opt} m}{1 + \lambda_{opt}} \neq 0 \text{ for } m \neq 0, \lambda_{opt} \neq 0$

3 (100pts) Weight Decay and Gradient Descent

$w_i(t) = w_i(t-1) \underbrace{-\eta \nabla_w \mathcal{E}_0}_{\text{ignore}} - 2\eta\lambda w_i(t-1) \Rightarrow w_i(t) = (1 - 2\eta\lambda)w_i(t-1)$. Iterating this equation we see that

$$w_i(t) = (1 - 2\eta\lambda)^t w_i(0)$$

which represents exponential decay provided that $|1 - 2\eta\lambda| < 1$ which requires $0 < \eta\lambda < 1$.

4 (250pts) Bootstrapping

(a) $P[x_i^j = x_\alpha] = \frac{1}{N}$ for any x_α . $P[x_i^j \neq x_\alpha] = 1 - \frac{1}{N}$ and so $P[k \text{ occurrences of } x_\alpha]$ is given by the binomial distribution

$$P[k] = \binom{N}{k} \left(\frac{1}{N}\right)^k \left(1 - \frac{1}{N}\right)^{N-k}$$

(b) (i) We consider a more general problem. Let $g(\cdot)$ be any function and let $f(S) = \frac{1}{N} \sum_{i=1}^N g(x_i)$. In our case, $g(x) = x^r$. Note that $E_B[g(x_j^i)] = \frac{1}{N} \sum_{k=1}^N g(x_k) = f(S)$ for every i, j . Thus,

$$E_B[f(S_i)] = E_B \left[\frac{1}{N} \sum_{j=1}^N g(x_j^i) \right] = \frac{1}{N} \sum_{j=1}^N E_B [g(x_j^i)] = \frac{1}{N} \sum_{j=1}^N f(S) = f(S)$$

$$\text{Thus, } \boxed{E_B[f(S_i)] = f(S) \Rightarrow E_{p(x)}[E_B[f(S_i)]] = E_{p(x)}[f(S)]}$$

(extra credit 50 pts) First note that

$$Var_{p(x)}[f(S)] = Var_{p(x)} \left[\frac{1}{N} \sum_{i=1}^N g(x_i) \right] = \frac{1}{N^2} \sum_{i=1}^N Var_{p(x)}[g(x_i)] = \frac{1}{N} Var_{p(x)}[g(x)]$$

We now consider $Var_B[f(S_i)]$.

$$Var_B[f(S_i)] = Var_B \left[\frac{1}{N} \sum_{j=1}^k g(x_j^i) \right] = \frac{1}{N} Var_B [g(x_j^i)] = \frac{1}{N} \left(\frac{1}{N} \sum_{k=1}^N g(x_k)^2 - \frac{1}{N^2} \sum_{i,j=1}^N g(x_j^i)g(x_k^i) \right)$$

Thus,

$$\begin{aligned} E_{p(x)} [Var_B[f(S_i)]] &= \frac{1}{N} \left(\frac{1}{N} \sum_{k=1}^N E_{p(x)}[g(x_k)^2] - \frac{1}{N^2} \sum_{i,j=1}^N E_{p(x)}[g(x_j^i)g(x_k^i)] \right) \\ &= \frac{1}{N} \left(E_{p(x)}[g(x)^2] - \frac{N(N-1)}{N^2} E_{p(x)}[g(x)]^2 - \frac{N}{N^2} E_{p(x)}[g(x)^2] \right) \\ &= \frac{N-1}{N^2} (E_{p(x)}[g(x)^2] - E_{p(x)}[g(x)]^2) = \boxed{\frac{N-1}{N} Var_{p(x)}[f(S)]} \end{aligned}$$

(ii) Let $z_i = \begin{cases} 1 & \text{if } x_i \notin S_k \\ 0 & \text{if } x_i \in S_k \end{cases}$ where S_k is a bootstrapped set, and let $z = \sum_{i=1}^N z_i$ be the number of missing elements. The number of missing elements is exactly the number of repetitions, thus we want $E_B[z]$.

But $E_B[z] = \sum_{i=1}^N E_B[z_i] = NP[z_i = 1]$. But $P[z_i = 1] = P[x_i \text{ appears 0 times}] = (1 - \frac{1}{N})^N$ by part (a) thus

$$\boxed{\text{Expected \# of repetitions} = E_B[z] = N \left(1 - \frac{1}{N}\right)^N \approx \frac{N}{e}}$$

The bootstrap fails because with respect to the function $f(S)$, the set S does not adequately represent the input distribution. In fact, this is so even as $N \rightarrow \infty$ so the bootstrap has no chance (unless $N = 1$).

5 (Optional Bonus 50pts) Cross Validation

(a) $E_{D_N}[e^{(i)}] = E_{D_{N-1}^{(i)}} E_{x_i}[e^{(i)}] = E_{D_{N-1}^{(i)}} [P[g_{D_{N-1}^{(i)}} \neq f]] = \boxed{\pi_{N-1}}$.

(b) $E_{D_N}[e] = \frac{1}{N} \sum_{i=1}^N E_{D_N}[e^{(i)}] = \frac{1}{N} \sum_{i=1}^N \pi_{N-1} = \boxed{\pi_{N-1}}$.

We now consider the variance of e . The variance of $e^{(i)}$ is given by $Var(e^{(i)}) = \pi^{(i)}(1 - \pi^{(i)})$.

(c) Assuming that the $e^{(i)}$'s are independent,

$$Var[e] = \frac{1}{N^2} \sum_{i=1}^N Var[e^{(i)}] = \frac{1}{N^2} \sum_{i=1}^N \pi^{(i)}(1 - \pi^{(i)}) = \frac{1}{N} \hat{\Pi} - \frac{1}{N^2} \sum_{i=1}^N \pi^{(i)^2}$$

but $\frac{1}{N} \sum_{i=1}^N \pi^{(i)^2} - \left(\frac{1}{N} \sum_{i=1}^N \pi^{(i)}\right)^2 \geq 0 \Rightarrow \frac{1}{N} \sum_{i=1}^N \pi^{(i)^2} \geq \hat{\Pi}^2$ hence

$$\boxed{Var[e] \leq \frac{1}{N} \hat{\Pi} - \frac{1}{N} \hat{\Pi}^2 = \frac{1}{N} \hat{\Pi}(1 - \hat{\Pi})}$$