

ASSIGNMENT 4, due Mon, Mar. 26

Discussions are encouraged but no written records can be taken away from a discussion. Books and notes can be consulted but not copied from. Homeworks are due in class or in my mail box by 4pm on the due date. For experimental problems, debugging help may be obtained but all code should be your own and all results reported should be from running your own code. In order to verify “curious” results, we may request you submit your code so DO NOT delete your code after you hand in a problem set.

1 (300pts) Computation of BIAS and VARIANCE

In this problem, we consider a simplified learning scenario. Assume that the input dimension is 1. Assume that the input variable x is uniformly distributed in the interval $[0, 1]$. The data set consists of 2 points $\{x_1, x_2\}$ and assume that the target function is $f(x) = x^2$. Thus, the full data set is $\{(x_1, x_1^2), (x_2, x_2^2)\}$. Assume that the learning model is the set of linear functions (not perceptrons) – we will be doing regression here. Thus the learning model consists of functions of the form $g(x) = ax + b$ where a, b are any constants. The learning algorithm is to find the constants that fit the 2 data points exactly. We have thus defined our learning system. Given any data set, the learning system will yield a (linear) function. We are interested in the test performance (R) of our learning system with respect to the squared error measure, the BIAS and the VAR. Remember the following definitions

$$R = E_{D_N} \left[\int_0^1 dx p(x) (f(x) - g(x, D_N))^2 \right] \quad \hat{g}(x) = E_{D_N} [g(x, D_N)]$$
$$BIAS = \int_0^1 dx p(x) (f(x) - \hat{g}(x))^2 \quad VAR = \int_0^1 dx p(x) E_{D_N} [(g(x, D_N) - \hat{g}(x))^2]$$

where in our case $p(x) = 1$.

- Describe an experiment that you could run to determine (numerically) $\hat{g}(x)$, R , $BIAS$, and VAR for this learning system.
- Run your experiment and report the results. compare R with $BIAS + VAR$. Provide a plot of your $\hat{g}(x)$ and $f(x)$ (on the same plot).
- (extra credit, 50pts) Compute analytically what the results should have been.

2 (350pts) Weight Decay Trades Bias for Variance

In this problem, we will compute the optimal weight decay parameter (λ) for a simple learning problem. Let the data set be a set of numbers drawn from a distribution with unknown mean m : $D_N = \{y_i\}_{i=1}^N$ where each $y_i = m + \epsilon_i$, where ϵ_i is noise. The task is to estimate the mean m from the $\{y_i\}$. We will call our estimate \hat{m} . Assume the noise to be zero mean and independent, with

$$E[\epsilon_i] = 0 \quad E[\epsilon_i \epsilon_j] = \sigma^2 \delta_{ij}$$

where $\delta_{ij} = 1$ if $i = j$, otherwise zero.

- (a) Assume each ϵ_i to have a Gaussian (Normal) distribution with mean 0 and variance σ^2 :

$$P(\epsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon^2}{2\sigma^2}}$$

Let $\mathcal{L} = \mathbf{R}$ be the learning model. A “function” in our learning model is just a number $w \in \mathbf{R}$. Our estimate \hat{m} will be the w that maximizes $Pr[m = w|D_N]$. Assume that $Pr[m = w]$ is uniform, independent of w . Then show that minimizing the error function $\mathcal{E}_0 = \frac{1}{N} \sum_{i=1}^N (w - y_i)^2$ with respect to w leads to the desired estimate \hat{m} . Call this estimate \hat{m}_0 . Show that $\hat{m}_0 = \frac{1}{N} \sum_{i=1}^N y_i$.

- (b) Now consider this same error function with a weight decay term added

$$\mathcal{E} = \mathcal{E}_0 + \lambda w^2$$

Minimizing this, one obtains a second estimate \hat{m} . Show that $\hat{m} = \hat{m}_0 / (1 + \lambda)$, hence the term weight decay for $\lambda > 0$.

- (c) Compute $\mathcal{R}(\lambda) = E_{\epsilon_1, \dots, \epsilon_N} [(\hat{m} - m)^2]$, the expected value of the squared deviation with respect to the noise. This is a function of λ .
- (d) Minimize this with respect to λ to obtain the optimal lambda (λ_{opt}), that value of λ that minimizes the expected test error. Obtain the expected squared deviation that results from using λ_{opt} and show that

$$\mathcal{R}(\lambda_{opt}) = \frac{\mathcal{R}(0)}{1 + \lambda_{opt}} < \mathcal{R}(0)$$

You have shown the hopefully unintuitive statement: if I give you N independent drawings from a distribution and ask you to estimate the mean, your criterion being squared deviation, then you should not use the sample mean! For example using the sample mean decreased by a factor of $1/(1 + \lambda_{opt})$ does better. The best UNBIASED estimator of the mean, however, is the sample mean.

- (e) Show that the estimate \hat{m} using λ_{opt} is statistically biased (i.e., $bias = E_{\epsilon_1, \dots, \epsilon_N} [m - \hat{m}] \neq 0$) and determine the value of the bias. What our weight decay estimator has done is trade off a little bit of bias for a more significant gain in the variance.

3 (100pts) Weight Decay and Gradient Descent

When using weight decay with gradient descent, one simply adds a term λw to the gradient (i.e., $\nabla_w \mathcal{E} = \nabla_w \mathcal{E}_0 + \lambda w$). If we ignore the $\nabla_w \mathcal{E}_0$ term, then show that the weight update would be $\Delta w_i = -\eta \lambda w_i$. Show that this leads to exponential decay of the weights for $\eta \lambda$ in a certain range, hence the term weight decay.

4 (250pts) Bootstrapping

We are given a sample $S = \{x_1, \dots, x_N\}$ where $x_i \in \mathbf{R}$ and x_i are drawn independently from $p(x)$. Let $f(S)$ be some function of S . Let S_1, \dots, S_K be K bootstrapped sets of size N . i.e., $S_i = \{x_1^i, \dots, x_N^i\}$ where each x_j^i is picked randomly with uniform probability from the original set S – sampling with replacement. Call this Bootstrapping distribution (the distribution from which each x_j^i is picked) B .

- (a) (Anatomy of the S_j) Show that the probability of exactly k occurrences of a single x_i in a given S_j is given by

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

- (b) The goal of the Bootstrap is to get a better estimate of $f(S)$ by taking an average over the Bootstrapped sets: $\frac{1}{K} \sum_{i=1}^K f(S_i)$. This can be the case only if the statistics of the $f(S_i)$ are similar to those of $f(S)$. We will consider two examples.

- (i) $f(S) = \frac{1}{N} \sum_{i=1}^N (x_i)^r$ for some r . Show that

$$E_B[f(S_i)] = f(S)$$

hence, $E_{p(x)}[E_B[f(S_i)]] = E_{p(x)}[f(S)]$. Thus the expectation is unaltered and hence the first order statistics of $f(S_i)$ mimic those of $f(S)$.

(extra credit 50 pts) Show that $E_{p(x)}Var_B[f(S_i)] = \frac{N-1}{N}Var_{p(x)}[f(S)]$, hence that the second order statistics are also almost preserved in the bootstrapped sets.

- (ii) $f(S) = \text{total\# repeated elements in } S$. By repeated elements is meant the number of times an element is repeated. For example the set $\{1, 1, 1, 1\}$ has 4 repeated elements (not 1, which is the number of elements that are repeated and not 5, the number of times the element “1” appears). As another example, the set $\{1, 1, 2, 2\}$ contains 3 repeated elements. For continuous distributions $p(x)$, the number of repetitions is 0 with probability 1. Show that for the bootstrapped sets however,

$$E_B[f(S_i)] = N \left(1 - \frac{1}{N}\right)^N \approx \frac{N}{e}, \text{ for } N \text{ large}$$

i.e., on average, the Bootstrapped sets S_i will each have N/e repeated elements as compared with S which will have no repeated elements with probability 1, thus the bootstrap has no chance. What went wrong?

5 (Optional Bonus 50pts) Cross Validation

We have a learning system that maps a data set of size N to a hypothesis function. $\mathcal{L} : D_N \rightarrow g_{D_N}$. Let $R(g_{D_N})$, the risk associated with g_{D_N} be the probability that $g \neq f$. Taking the expectation with respect to the data set, we define

$$\pi_N = E_{D_N}[Pr[g_{D_N} \neq f]]$$

This represents the expected error when training on N data points. Denote by $D_{N-1}^{(i)}$, the data set D_N excepting for the i^{th} data point, and let $\mathcal{L} : D_{N-1}^{(i)} \rightarrow g_{N-1}^{(i)}$. Now compute the error of $g_{N-1}^{(i)}$ on the i^{th} data point (the one that was left out).

$$e^{(i)} = \begin{cases} 0, & g_{N-1}^{(i)}(x_i) = y_i \\ 1, & g_{N-1}^{(i)}(x_i) \neq y_i \end{cases}$$

$e^{(i)}$ is an estimate of the test error that results from training with $D_{N-1}^{(i)}$. Let $\pi^{(i)} = E_{x_i}[e^{(i)}]$.

- (a) Show that $E_{D_N}[e^{(i)}] = \pi_{N-1}$. In other words, each $e^{(i)}$ is an unbiased estimate of the expected test error when training on a data set of size $N - 1$.
- (b) Let $e = \frac{1}{N} \sum_{i=1}^N e^{(i)}$. Show that $E_{D_N}[e] = \pi_{N-1}$. Hence, e is also an unbiased estimate of the expected test error when training on a data set of size $N - 1$.

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, show that

$$Var(e) \leq \frac{1}{N} \hat{\Pi}(1 - \hat{\Pi})$$

where $\hat{\Pi} = \frac{1}{N} \sum_{i=1}^N \pi^{(i)}$. Thus we see that the variance of e would decrease like $1/N$ provided we had independence. Ofcourse, in reality, we don't have full independence so we don't gain the benefit of the full $1/N$.