Mouselinos Spyridon HW4

Exer Size 1

Consider the Regression Problem $y = g(x) + n$

Min MSE Estimate $E[y|x]$ / Estimator $f(x;D)$

(a) The quantity $E_D[(f(x;D) - E[y|x])^2]$ is the MSE of our estimator $f(x;D)$. It can be broken down into two parts:

$$\text{Bias}^2: \left(E_D[f(x;D)] - E[y|x]\right)^2$$

$$+$$

$$\text{Variance}: E_D[f(x;D) - E_D[f(x;D)]^2]$$

$$=$$

$$MSE$$

Now in case that we have a finite number of training points/samples, there is a tradeoff between the two terms as they can't be reduced simultaneously. As the bias decreases, meaning we opt for a more complex model, the variance increases -> meaning inability to generalize between samples and on new data.

To become zero we must have an unbiased estimator thus having exactly the same complexity and form as the data generation process $E_D[f(x;D)] = E[y|x]$, as well as an infinite number of training points, that will force our estimator not to fluctuate between samples.

(b) This can't be achieved in practice for 2 reasons:

·) We can't perfectly guess the underlying data generation mechanism. If we knew, why bother estimating it

·) We can't have -practically- an infinite number of samples for our Dataset

Exersize 2

Regression task: $y = g(x) + n$

f_θ -> estimator of $g(x)$, parametrized by $\vec{\theta}$.

Tr -> Train Set.

Te -> Test Set.

(a) A large error value in Tr may indicate $\boxed{\text{large bias}}$ in our estimator. That means that we have chosen an f_θ of the wrong family of equations, or an f_θ with less expressivity in terms of parameters that required.

(b) A large error value in Te on the test set may derive either from (a), meaning a poor choice of model would perform bad - (even worse in most cases) - on unseen data. If our estimator performed good in Tr but only bad in Te then we might suffer from overfitting, meaning our estimator has $\boxed{\text{high variance}}$. That could mislead us with a good fit on Tr. The estimator has too may parameters and is over complex for the required task.

Exersize 2

(c) A small error value in Tr may derive from (b), meaning we have created such a complex model that can fully capture perfectly the paints in our dataset - almost zero bias - but high variance, or we simply chose a got model to explain the nature of our data generation, with the form of f_θ beign close to the $E[y|x]$.

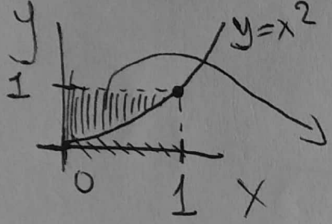
(d) A small error value in Te is a very good sign as it generally means the model that was chosen was a good balance between the bias-variance tradeoff, and could both fit well on the training data, as well as generalize well on unseen data.

Exersize 3

Let's consider a regression task $y = g(x) + \eta$.
Where x, y are RV's with joint pdf:

$$p(x,y) = \frac{3}{2}, \quad x \in (0,1) \text{ and } y \in (x^2, 1)$$

(a) First lets plot and verify that $p(x,y)$ is a pdf.



The area of interest.

According to the Kolmogorov axioms we need:

$$\int_A p(x,y)\,dA = 1 \Rightarrow \int_0^1 \left[\int_{x^2}^1 p(x,y)\,dy\right] dx = 1 \Rightarrow$$

$$\int_0^1 \left[\int_{x^2}^1 3/2\,dy\right] dx = 1 \Rightarrow \frac{3}{2}\int_0^1 [y]_{x^2}^1\,dx = 1 \Rightarrow$$

$$\Rightarrow \int_0^1 1-x^2\,dx = \frac{2}{3} \Rightarrow \int_0^1 \left(x-\frac{x^3}{3}\right)' dx = \frac{2}{3} \Rightarrow \left[x-\frac{x^3}{3}\right]_0^1 = \frac{2}{3}$$

$$\Rightarrow \left[1 - 1/3 - 0 + 0\right] = \frac{2}{3} \Rightarrow \frac{2}{3} = \frac{2}{3}$$ OK

(b) Compute the marginal pdf of x, $p_X(x)$.

We know that:

$$p_x(x) = \int_{x^2}^{1} p(x,y)\,dy = \int_{x^2}^{1} 3/2\,dy = 3/2\,[1-x^2]$$

$$= \frac{3}{2} - \frac{3}{2}x^2.$$

(c) The conditional probability of y given x is:

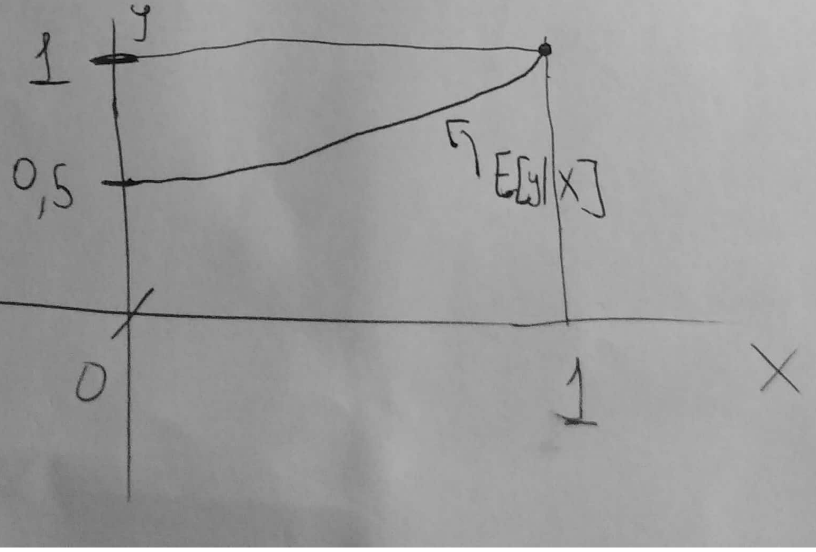
$$p(y|x) = \frac{p(x,y)}{p_x(x)} = \frac{3/2}{3/2\,[1-x^2]} = \frac{1}{1-x^2}.$$

a function of x, as expected.

(d) $$E[y|x] = \int_{x^2}^{1} y\,p(y|x)\,dy = \int_{x^2}^{1} y\,\frac{1}{1-x^2}\,dy = \frac{1}{1-x^2}\int_{x^2}^{1}\left(\frac{y^2}{2}\right)' dy$$

$$= \frac{1}{1-x^2}\left[\frac{1}{2} - \frac{x^4}{2}\right] = \frac{-1}{x^2-1}\left[\frac{-(x^4-1)}{2}\right] = \frac{x^4-1}{2(x^2-1)} = \frac{(\cancel{x^2-1})(x^2+1)}{2(\cancel{x^2-1})}$$

$$= \frac{x^2+1}{2} \quad \text{with } x \in (0,1)$$



Comments on Ex.4

(a) We used the formula $E[y|x] = m_y + a\frac{s_y}{s_x}(x - m_x)$ assuming that our data follow the $N(m_{y|x}, S_{y|x})$ distribution and showed that is a straight line.

(b) (c) (d) We used the LS estimator that we created in the previous HW (HW3) in order to perform a LS on data. Our datasets consisted of 50 points and were 100 in number. From the plots we see a great variation in the estimated parameters that is eliminated when we get the mean of them. By eliminating the between-fit variance we get a near-perfect estimate.

(e) We re-did the previous steps but now our datasets, although same in number, consisted of 1000 points. That lead to far better estimates with both lower bias and variance. We once again estimated the average of them in order to eliminate the between-fit variance that lead to an estimator indistinguishable from the optimal one

(e) What this exersize shows us is that the MSE = Variance + Bias^2 can be partially eliminated by sampling over our estimators but is greatly reduced <u>per-estimator</u> when the number of points increases.
With an inf number of samples we could theoretically have a perfect fit both in terms of bias as well as in variance.

Comments on Ex 5

(a) After solving exensize 3 we found that under MSE loss the best estiomate is given as $E[y|x] = \frac{x^2+1}{2}$. We plotted in red the area our points would appear under their pdf and we plotted them in blue. After, for each point we plotted in green $\boxed{x}$'s its estimate under ex3. Meaning for each x we plotted $\frac{x^2+1}{2}$.

(b)(c) Under the erroneous assumption of a ~~Gauss~~ Normal distribution we use the same $E[y|x] = my + a\frac{Sy}{Sx}(x-mx)$ estimator for each point and re-plot everything, where now our erroneas estimates are in orange stars "*".

We see that we are not really off in this scenario by observing the MSE of the best and this solution. Howerer, it is mathematically proven to be a worse solution.