Problem 1 Amin Asadi

a) In the first one the train-test partition is 50% train & 50% test
In the second one the train portion is 95% and the test
portion is 5%. So because in the second one the train
set is larger than first one, the second model is more generalized
and it's parameter's estimation have less variance. However
since the test set is smaller in the second one, the performance
result's variance is larger in the second one.
The reason for the 90% accuracy being more than the 80%
accuracy might be the more generalized model in the second one,
but the accuracy of the second model may be very different
on another test set (because as we noted above, it has larger
variance) and hence the 90% accuracy is not reliable.
Also increasing the trainset size from 200 to 380 has not effected
to much in the performance in the second model. It could have been
better to increase trainset size for example by 100 and
instead have a bit larger train set in the second model

b) Statistical Inference = Probability^{-1}

+ In probability we are concerned about the properties of
data from a specified probability distribution
For example if we know that Random Variables $X_1, \ldots, X_5$ have a
standard Gaussian distribution and they are iid, what is
the probability of X being more than 0.5 ($P(X>0.5)$)
+ In Statistical Inference we are concerned about finding
the properties of the distribution of a specified set of data.
For example suppose that we know $X_1 = 1$ and $X_2 = 2$, ...
and $X_5 = 3$. We are concerned about finding parameter (mean) θ
of the distribution $N(\theta, 1)$ such the $X_1, \ldots, X_5 \overset{iid}{\sim} N(\theta, 1)$
We use Inference mainly in Hypothesis testing, Estimation and
Confidence Intervals.

Amin Asadi — Machine Learning Homework #1 — Fall 2020

Problem 1:

a) Suppose that we have a function F which is dependent on one or more variables. The goal of gradient descent is to find the local minimum of F. This algorithm is based on the fact that, at any point x that F is differentiable on a neighbourhood of x, the fastest decrease in F, occurs when we go from x proportional to negative of gradient of F at point x. We can find the local minimum of F by iteratively taking such steps.

Here $F = J(\theta) = J(w;b) = \frac{1}{2} \sum_{i=1}^{q} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

$$\boxed{\frac{\partial J}{\partial b}} = \frac{1}{2} \sum_{i=1}^{q} 2\left(h_\theta(x^{(i)}) - y^{(i)}\right) \left(\frac{\partial (w^T x^{(i)} + b)}{\partial b} \right) \left(1 - \tanh^2(w^T x^{(i)} - b)\right) = \sum_{i=1}^{q} \left(h_\theta(x^{(i)}) - y^{(i)}\right)\left(1 - \tanh^2(w^T x^{(i)} + b)\right)$$

$$\boxed{\frac{\partial J}{\partial w_j}} = \frac{1}{2} \sum_{i=1}^{q} 2\left(h_\theta(x^{(i)}) - y^{(i)}\right) \left(\frac{\partial h_\theta(x^{(i)})}{\partial w} \right) = \sum_{i=1}^{q} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)} \left(1 - \tanh^2(w^T x^{(i)} - b)\right)$$

$w_j \leftarrow$ $W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$ $x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}$ $\rightarrow x_j^{(i)}$

b) If learning Rate $= \alpha$:

update rules: $b := b - \alpha \frac{\partial J}{\partial b}$, $w_j := w_j - \alpha \frac{\partial J}{\partial w_j}$

If the learning rate is small then we will take small steps towards the minimum but we are sure that we keep getting closer and closer to that point (It will converge to local minimum) On the other hand, if the learning rate is too high, we will take bigger steps and thus we will get to the minimum in less steps but there is high chance that we overshoot the local minimum and repeatedly fall into other side of the point and hence do not converge. So the learning rate should be tuned as a hyper parameter.

furthor explanation of the derivative in part a:

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{2} \sum_{i=1}^{q} 2\left(h_\theta(x^{(i)}) - y^{(i)}\right) \frac{\partial}{\partial \theta_j}\left(h_\theta(x^{(i)}) - y^{(i)}\right)$$

$$= \frac{1}{2} \sum_{i=1}^{q} 2\left(h_\theta(x^{(i)}) - y^{(i)}\right) \frac{\partial}{\partial \theta_j}\left(\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \cdots + \theta_m x_m^{(i)} + b - y^{(i)}\right) \times \frac{d}{dx}(\tanh(x))\Big|_{w^T x^{(i)} + b}$$

$$= \sum_{i=1}^{q} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)} \left(1 - \tanh^2(w^T x^{(i)} + b)\right)$$

Problem 3 :

We prove that there is a unique polynomial of degree K or less that passes through these K+1 points :

let $p(x) = a_K x^K + a_{K-1} x^{K-1} + \dots + a_1 x + a_0$ be a polinomial $(a_i \geq 0)$

If we substitute the K+1 points in $p(x)$ we will have a system of K+1 linear equations:

K+1 points : $(x_0, y_0), \dots, (x_K, y_K)$

$$a_K x_0^K + a_{K-1} x_0^{K-1} + \dots + a_0 = y_0$$

$$\vdots$$

$$a_K x_K^K + a_{K-1} x_{K-1}^{K-1} + \dots + a_0 = y_K$$

$$\Downarrow$$

$$\underbrace{\begin{bmatrix} x_0^K & x_0^{K-1} & \dots & x_0 & 1 \\ x_1^K & x_1^{K-1} & \dots & x_1 & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ x_K^K & x_{K-1}^{K-1} & & x_K & 1 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} a_K \\ a_{K-1} \\ \vdots \\ a_0 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_K \end{bmatrix}}_{Y}$$

X is a Vandermont Matrix with $\det(X) = \prod_{1 \leq i \leq j \leq K+1} (x_j - x_i)$

because K+1 points have distinct x $\rightarrow \det(X) \neq 0$

$\rightarrow A = X^{-1} Y \longrightarrow$ So $P(x)$ is unique.

Obviously if p had a degree more than K, the number of variables would be more than number of equations and hence the system would have more than one sulutions $\rightarrow$ more than one polynomials

Problem 4:

suppose $\leftarrow$ The X_i are fixed

$Y = b_0 + b_1 x$
$Var(Y_i) = \sigma^2$

$$\beta_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})Y_i - \bar{Y}\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$

$$= \frac{\sum(x_i - \bar{x})Y_i - \bar{Y}\sum x_i + \bar{Y} \cdot n \cdot \sum \frac{x_i}{n}}{\sum(x_i - \bar{x})^2} = \sum \overbrace{\frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}}^{K_i} Y_i = \sum K_i Y_i \quad (1)$$

$$\sum_i K_i = \sum_i \frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2} = \frac{1}{\sum_i (x_i - \bar{x})^2} \underbrace{\cancel{\sum_i (x_i - \bar{x})}}_{0} = 0 \quad (2)$$

$$\sum_i K_i^2 = \sum_i \left(\frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2}\right)^2 = \frac{\sum(x_i - \bar{x})^2}{\left(\sum(x_i - \bar{x})^2\right)^2} = \frac{1}{\sum(x_i - \bar{x})^2} \quad (3)$$

$$\sum_i K_i x_i = \frac{\sum_i x_i^2 - x_i \bar{x}}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i x_i^2 - 2x_i\bar{x} + \bar{x}^2 + x_i\bar{x} - \bar{x}^2}{\sum_i (x_i - \bar{x})^2}$$

$$= \frac{\sum_i (x_i - \bar{x})^2 + \bar{x} \underbrace{\cancel{\sum_i (x_i - \bar{x})}}_{0}}{\sum_i (x_i - \bar{x})^2} = 1 \quad (4)$$

$$E(\beta_1) = E(\sum K_i Y_i) = \sum K_i E(Y_i) = \sum K_i (\beta_0 + \beta_1 x_i)$$

$$\overset{(2)(4)}{=} \beta_0 \underbrace{\cancel{\sum K_i}}_{0} + \beta_1 \underbrace{\cancel{\sum K_i x_i}}_{1} = b_1 \quad (5)$$

$$Var(\beta_1) = Var(\sum K_i Y_i) = \sum K_i^2 Var(Y_i) + \sum_i \sum_{j \neq i} K_i K_j \underbrace{\cancel{Cov(Y_i, Y_j)}}_{0} = Var(Y_i) \sum K_i^2$$

$$\overset{(3)}{=} \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

$$Cov(\beta_0, \beta_1) = E\left[(\beta_0 - E(\beta_0))(\beta_1 - E(\beta_1))\right] \quad (6) \quad \text{By Definition}$$

$$E(\beta_0) = \bar{Y} - E(\beta_1)\bar{x} \overset{(5)}{=} \bar{Y} - b_1 \bar{x} \quad (7)$$

$$(7) \rightarrow \beta_0 - E(\beta_0) = \beta_0 - (\bar{Y} - \beta_1 \bar{x}) = \bar{Y} - \beta_1 x - \bar{Y} + \beta_1 \bar{x}$$

$$= -\bar{x}(\beta_1 - b_1) \quad (8)$$

$$(7),(8) \Rightarrow Cov(\beta_0, \beta_1) = E\left[(-\bar{x}(\beta_1 - b_1))(\beta_1 - b_1)\right]$$

$$= E(-\bar{x}(\beta_1 - b_1)^2) = -\bar{x}\, E(\beta_1 - b_1)^2$$

$$= -\bar{x}\, Var(\beta_1) = -\bar{x}\frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

$\beta_0, \beta_1 \rightarrow$ independent $\rightarrow Cov(\beta_0, \beta_1) = 0 \rightarrow -\bar{x}\frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} = 0$

$\rightarrow \begin{cases} \bar{x} = 0 \rightarrow \\ \sigma^2 = 0 \rightarrow \end{cases}$

$\bar{x} = 0 \rightarrow$ میانگین داده ها صفر باشد

$\sigma^2 = 0 \rightarrow$ واریانس خروجی ها صفر باشد ← خروجی ها برابر باشند

Problem 5:

a) $\sigma^2 = Var(Y) = \dfrac{\sum_{i=1}^{n}(Y_i - \overline{Y}_i)}{n-1}$

در همه واریانس ها درجه آزادی را n-1 فرض کرده ایم

$\overline{Y}_i = \dfrac{\sum_{i=1}^{8} Y_i}{n} = \dfrac{40+41+43+42+44+42+43+42}{8} = 42.125$

$\rightarrow \sigma^2 = \dfrac{\sum_{i=1}^{8}(Y_i - 42.125)}{7} \approx 1.55$

$\overline{x}_i = \dfrac{0.5+1+\dots+4}{8} = 1.5$

$\beta_1 = \dfrac{\sum_{i=1}^{8}(Y_i - \overline{Y}_i)(x_i - \overline{x}_i)}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \dfrac{\sum_{i=1}^{8}(Y_i - 42.125)(x_i - 1.5)}{\sum_{i=1}^{8}(x_i - 1.5)^2} \approx 0.54$

$\beta_0 = \overline{Y} - \beta_1 \overline{x} = 42.125 - 0.54 \times 1.5 \approx 40.89$

Based on (1) in problem 4 solution: $\begin{cases} \beta_1 = \sum K_i Y_i \\ K_i = \dfrac{x_i - \overline{x}}{\sum(x_i - \overline{x})^2} \end{cases}$

By variance sum law: $Var(\beta_1) = \sum_{i=1}^{n} K_i^2 Var(Y_i) + \sum_i \sum_{j \neq i} K_i K_j \cancel{Cov(Y_i, Y_j)}$

$\rightarrow Var(\beta_1) = \sum_{i=1}^{n} K_i^2 \overset{\sigma^2}{Var(Y_i)} = \sigma^2 \sum_{i=1}^{n} K_i^2 = \dfrac{\sigma^2}{\sum(x_i - \overline{x})^2}$

independency $\rightarrow 0$

$= \dfrac{1.55}{\sum_{i=1}^{8}(x_i - 1.5)^2} = \dfrac{1.55}{10.5} \approx 0.147$

$Var(\sum \frac{Y_i}{n}) = \frac{1}{n^2} \sum Var(Y) = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}$

$\beta_0 = \overline{Y} - \beta_1 \overline{x} \rightarrow Var(\beta_0) = Var(\overline{Y}) + Var(\beta_1 \overline{x}) - 2Cov(\overline{Y}, \overline{x}\beta_1)$

$= \dfrac{\sigma^2}{n} + \overline{x}^2 Var(\beta_1) - 2\overline{x}\, Cov(\overline{Y}, \beta_1)$

$Cov(\overline{Y}, \beta_1) = E[(\overline{Y} - \cancel{E(\overline{Y})})(\beta_1 - E(\beta_1))] = 0$

$\rightarrow Var(\beta_0) = \dfrac{\sigma^2}{n} + \overline{x}^2 Var(\beta_1) = \dfrac{\sigma^2}{n} + \overline{x}^2 \dfrac{\sigma^2}{\sum(x_i - \overline{x})^2} = \sigma^2\left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{\sum(x_i - \overline{x})^2}\right)$

$\rightarrow Var(\beta_0) = \dfrac{1.55}{8} + (1.5)^2(0.147) = 0.52$

b) $Cov(\beta_0, B_1) = \dfrac{-\bar{x}\,\sigma^2}{\sum(x_i-\bar{x})^2} = \dfrac{-1.5 \times 1.55}{10.5} \approx -0.22$

$Cor(B_0, B_1) = \dfrac{Cov(\beta_0, \beta_1)}{Var(\beta_0)\,Var(\beta_1)} = \dfrac{-0.22}{0.52 \times 0.147} = -2.87$

Problem 8:

Suppose for example that we have built a classifier which predicts if someone has cancer or not. We have tested the classifier with 1000 samples and the confusion matrix is:

Actual \ predict	Positive	Negative
Positive	1	1
Negative	0	998

Accuracy: This metric measures that how much the model was good in predicting the correct class. Accuray $= \frac{\# \text{correct predictions}}{\# \text{total predictions}}$

Here accuracy $= \frac{1+998}{1+1+0+998} = \frac{999}{1000} = 99.9\%$ which is very high

Precesion: This metric measures that what fraction of the positive predictions are actually positive.

$$\text{precision} = \frac{\# \text{true positive}}{\# \text{total positive predictions}} = \frac{1}{0+1} = 100\%$$

Recall: This metric measures that what fraction of actual positives were predicted correctly, i.e. positive.

$$\text{recall} = \frac{\# \text{true positive}}{\# \text{total actual positives}} = \frac{1}{1+1} = 50\%$$

In this example although accuracy is very high (99.9%) but the importing thing is that samples which actually have cancer (actual positives) must not be predicted negative, othewise their lives are put in danger. If an actual negative is predicted positive is bad but it does not have as much cost as low recall. So in our example high recall is the most important.

When we want to design a spam email classifier, precision is the most important metric because we do not want an important email to be predicted spam. But if an spam email is predicted non-spam it does not have a high cost. In other examples where the general correctness of classification is important and false predictions only are not of a too much cost, the accuracy is the most important metric.