Neural Networks
Assignment 5

Group Members:

Ehtisham Ali - 2567631
Anastasiia Kalmykova - 2567127
Hafeez Ullah Milan - 2572872

Exercise 5.1.

(a)

(Maximum Likelihood Estimation) MLE.

$$P(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \quad - (i)$$

(i) is the density function for one occurance, for the whole training date it becomes.

$$L(\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)$$

taking ln Naturae log on both sides

$$LL(\mu) = \sum_{i=1}^{N}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(x_i-\mu)^2 + \log(e)\right] = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N - \sum_{i=1}^{N}\left[\frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

$$= N\log(1) - N\log(2\pi\sigma^2)^{\frac{1}{2}} - \sum_{i=1}^{N}\left[\frac{1}{2\sigma^2}(x_i-\mu)^2\right] = \frac{\partial LL(\mu)}{\partial \mu} = 0 - 0 -$$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left[2(x_i-\mu)\cdot(-1)\right]$$

$$= \frac{\cancel{2}\cdot 1}{\cancel{2}\sigma^2}\sum_{i=1}^{N}(x_i-\mu) = 0 \qquad \frac{\sum_{i=1}^{N}(x_i-\mu)}{\sigma^2} = 0$$

$$\sum_{i=1}^{N}(x_i-\mu) = 0$$

$$\sum_{i=1}^{N} x_i - \sum_{i=1}^{N}\mu = 0.$$

$$\sum_{i=1}^{N}\cdot\mu = \sum_{i=1}^{N} x_i$$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N}\cdot 1}$$

$$\boxed{\mu = \frac{\sum_{i=1}^{N}(x_i)}{N}}$$

As we know $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$, so μ is the mean in $P(x_i, \mu, \sigma^2)$ of univariate Gaussian distribution.

a (ii) MLE for variance (σ^2).

$$P(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(x-\mu)^2\right) \quad -(1)$$

Density function for the whole training data.

$$L(\sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(\frac{-1}{2\sigma^2}(x_i-\mu)^2\right)}$$

↳ taking natural log on both sides.

$$LL(\mu) = \sum_{i=1}^{N}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(x_i-\mu)^2 + \log(e)\right]$$

$$= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N - \sum_{i=1}^{N}\left[\frac{1}{2\sigma^2}(x_i-\mu)^2\right].$$

$$= N\log(1) - N\log(2\pi\sigma^2)^{1/2} - \sum_{i=1}^{N}\left[\frac{1}{2\sigma^2}(x_i-\mu)^2\right].$$

taking derivative w.r.t σ^2 and set to 0.

$$\frac{\partial}{\partial\sigma^2}\left[-\frac{N}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{N}\left[\frac{1}{2\sigma^2}(x_i-\mu)^2\right]\right] = 0$$

$$= \frac{-N \times \cancel{2\pi}}{2 \cdot \cancel{2\pi}\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}(x_i-\mu)^2 = 0.$$

$$\frac{\partial}{\partial\sigma^2}\left(\frac{1}{\sigma^2}\right) = \boxed{-\frac{1}{\sigma^4}}$$

$$\frac{+N}{2\sigma^2} = \frac{1}{2\sigma^4} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$\frac{+2\sigma^4 N}{2\sigma^2} = \sum_{i=1}^{N} (x_i - \mu)^2$$

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}.}$$

σ^2 is the variance in $p(x; \mu, \sigma^2)$ of univariate Gaussian distribution.

(b)

Minimum Likelihood Estimation for Linear Regression.

Assumption in linear Regression:

The independent variable $\vec{y}$ is assumed to be in a normal distribution.

In linear regression we take the model that we need to find, as the mean of the y's normal distribution.

So mean value would be $\boxed{w^T x_i \text{ for } y_i}$

Hence $P(y_i | x_i)$ is equivalent to normal distribution probability density function.

$$P(y_1, y_2 \ldots\ldots y_n | x_1, x_2 \ldots x_n) = \prod_{i=1}^{N} P(y_i | x_i)$$

as y in normal distribution.

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - w^T x_i)^2\right)$$

as $w^T x_i = x^T w_i$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i - x_i^T w)^2}{2\sigma^2}\right)$$

(b) Minimum Likelihood Estimation for Linear Regression.

Assumption in linear Regression.

The independent variable $\vec{y}$ is assumed to be in a normal distribution.

In linear regression we take the model that we need to find, as the mean of the y's normal distribution.

So mean value would be $\boxed{w^T x_i \text{ for } y_i}$

Hence $P(y_i | x_i)$ is equivalent to normal distribution probability density function.

$$P(y_1, y_2 \dots y_n | x_1, x_2 \dots x_n) = \prod_{i=1}^{N} P(y_i | x_i)$$

as y in normal distribution.

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - w^T x_i)^2\right)$$

as $w^T x_i = x^T w_i$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i - x_i^T w)^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp^{-}\left(\frac{\sum_{i=1}^{N}(y_i - x_i^T w)^2}{2\sigma^2}\right)$$

(2)

we know $(y_i - x_i^T w)^2 = (y_i - x_i^T w)^T (y_i - x_i^T w)$.

for whole traing dat $\sum_{i=1}^{N} x_i^T w = XW$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp^{-} \frac{(Y - XW)^T (Y - XW)}{2\sigma^2}$$

→ Now take natural log on both sides.

$$\ln(L(w)) = N \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(Y - XW)^T (Y - XW)}{2\sigma^2}$$

$$= N\log(1) - \frac{N}{2}\log(2\pi\sigma^2) - \frac{(Y - XW)^T (Y - XW)}{2\sigma^2}$$

= take derivatie w.r.t w on both sides and set $= 0$.

$$\frac{\partial LL(w)}{\partial w} = -\frac{1}{2\sigma^2} \frac{\partial}{\partial w} (Y^T - W^T X^T)(Y - XW)$$

$$= -\frac{1}{2\sigma^2} \frac{\partial}{\partial w} (Y^T Y - Y^T XW - W^T X^T Y + W^T X^T X W)$$

$$= -\frac{1}{2\sigma^2} \frac{\partial}{\partial w} (Y^2 - 2W^T X^T Y + W^T X^2 W)$$

$$= -\frac{1}{2\sigma^2} (0 - 2X^T Y + 2X^T X W)$$

optimal value of w when $\frac{\partial LL(w)}{\partial w} = 0$.

$$\frac{-1}{2\sigma^2}\left(-2X^TY + 2X^TXW\right) = 0.$$

$$-2X^TY + 2X^TXW = 0.$$

$$\cancel{2}X^TY = \cancel{2}X^TXW$$

$$\frac{X^TY}{X^TX} = W$$

$$\boxed{W = (X^TX)^{-1}X^TY} \quad (i)$$

eq (i) is the MLE for weights in case of Linear regression

↳ find weights optimal that minimize MSE.

$$MSE = \frac{1}{m}\|y - \hat{y}\|_2^2.$$

Set gradient to 0.

$$\nabla_w MSE = \nabla_w \frac{1}{m}\|y - \hat{y}\|_2^2$$

which comes out to equivalalent to (i).

$\boxed{\text{chapter 5 : slide 14}}$ reference.

Hence MLE of w $P(y_i|x_i)$ is equivalent to minimizing the MSE for w.

Holdout Method:

In holdout method we split the dataset into train and test sets. The training set is used to train the model and the test set is used to see how accurate the model performs on unseen data. Usually this split is of 80% train and 20% test.

Cross-Validation:

Cross-validation or k-fold cross-validation is when the dataset is randomly split into 'k' groups. This method is repeated k times and in each iteration one of the group is used to as a test set and the rest are used as a training set.

Why would one need cross validation instead of the holdout method

Holdout method is good to use when we have a really large dataset. While in the case of small dataset cross validation is needed with this method we can train on a large amount of data and still have all the data available for evaluation. Cross-validation on small data set reduce the variance in the model. Let's say we are using 5-fold CV and we end up 5 different models, we can select the model whose MSE is closest to

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$

Cross-validation is also needed in the scenario when we have a sparse dataset. In case of holdout method Machine Learning algorithm can suffer from the bias or variance of the chosen training set. The holdout estimate of the error would be misleading in this case, therefore we will be needing cross-validation.

Cross-Validation is also needed to complete feature selection for a given algorithm and tune its hyper-parameters. In this scenario cross-validation helps to reduce the amount of bias that enters the process due to the choices of parameters and CV allows to use a large amount of data to test those choices.

Exercise 5.2.

(b) · choices for the order of the polynomial.
$\{1, 5, 9\}$.

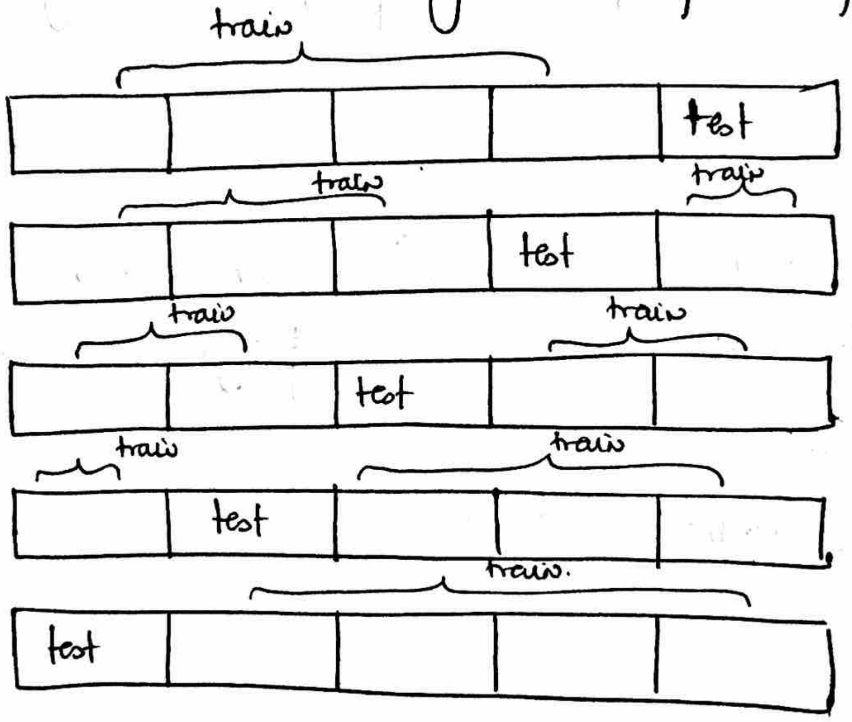
Hyper parameter selection.

We are going to use 5-fold cross validation.

Step-1:

Randomly divide the data set into 5-folds.

In $K=5$ repetitions, we select each single fold for testing and the remaing $K-1$ folds for training.

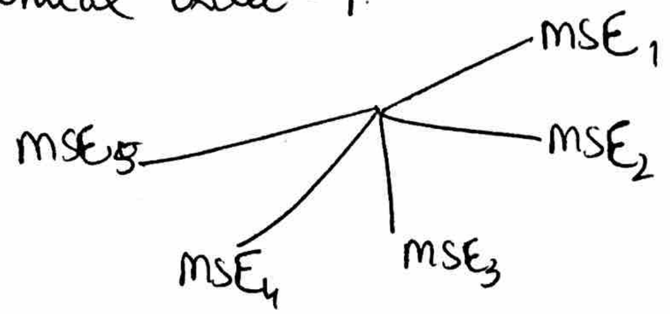


Step-2:

We will repeat this for each choice of polynomial order $= \{1, 5, 9\}$.

Hence for each choice of polynomial order we will get 5 different models.

Polynomial order - 1.



we get 5 different models with 5 MSEs.

We will compute the single final score

$$CV_5 = \frac{1}{5} \sum_{i=1}^{5} MSE_i \quad \text{--- (ii)}$$

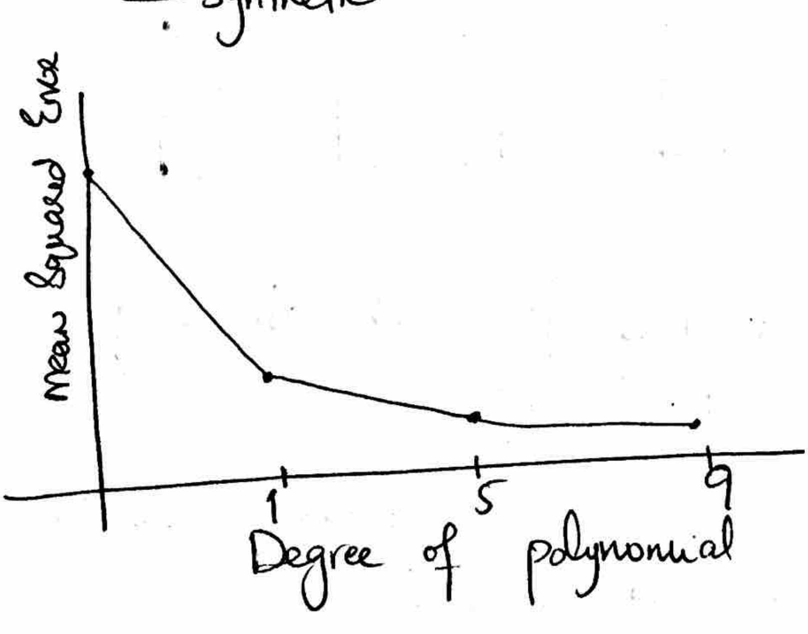
repeat the above step with polynomial degree 5 and 9.

<u>Step:3</u>

Now we get three CV_5 mean errors which we can use for the selection of polynomial degree.

we can plot the results.

— Synthetic data.



Exercise 5.3.

Logistic Regression.

(Multinomial) logistic regression is Generalized linear model (GLM) procedure, it uses the same basic formula of linear regression but instead of continous Y, it regresses the probability of a categorical outcome.

Multinomial logistic regression is the regression analysis to conduct when the dependent variable is nominal with more than two levels.

for-example.
Binary logistic regression assumes that the dependent variable is a stochastic event. The outcome is the probability with a density function to which class it belongs.

We use the linear regression when the output or dependent variable is a continous or discrete.
for categorical outcome we use logistic, regression.

I would selet the polynomial of degree 5 let's say. ②

Step: 4:

Now in polynoomial of degree 5, I have 5 models because of 5-fold CV.

I would select the model whose MSE is closest to $CV_5 = \frac{1}{5} \sum_{i=1}^{5} MSE_i$ to reduce to variance.