CS5691: Pattern Recognition and Machine Learning
Assignment #1

Topics: Regression, Classification, Density Estimation **Deadline:** 04 Oct 2021, 11:55 PM

Teammate 1: (Keyur Raval) **Roll number:** CS21M029
Teammate 2: (Makwana Rutik M.) **Roll number:** CS21M055

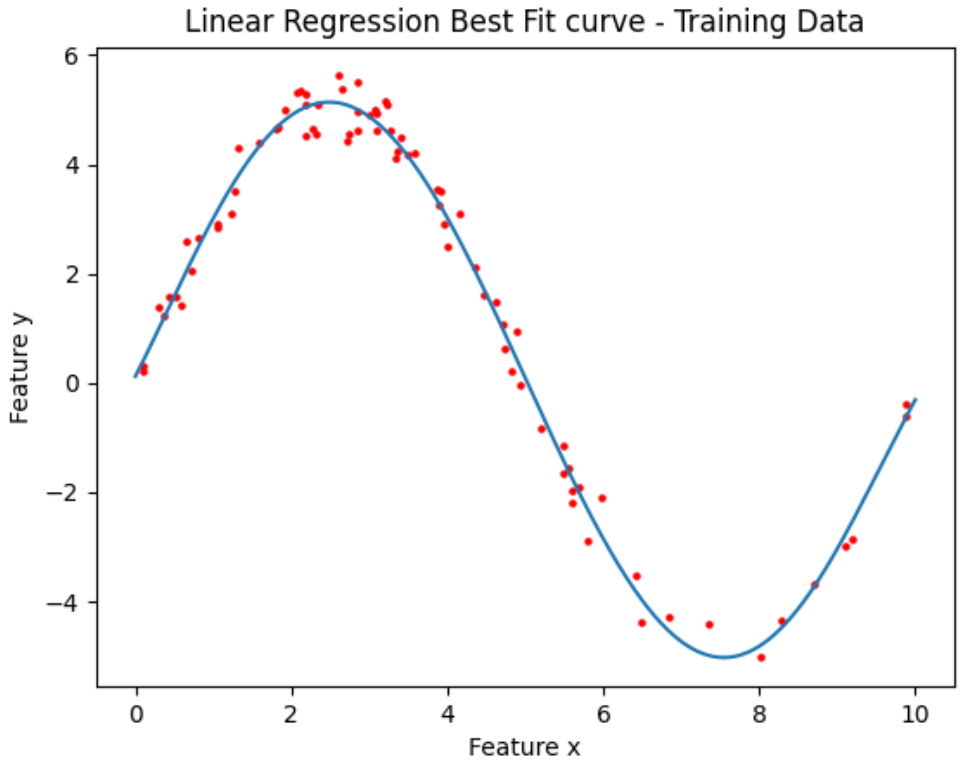
- This assignment has to be completed in teams of 2. Collaborations outside the team are strictly prohibited.
- Be precise with your explanations. Unnecessary verbosity will be penalized.
- Check the Moodle discussion forums regularly for updates regarding the assignment.
- Type your solutions in the provided LaTeX template file.
- For coding questions you will be required to upload the code in a zipped file to Moodle as well as embed the result figures in your LaTeX solutions.
- Attach a **README** with your code submission which gives a brief overview of your approach and a single command-line instruction for each question to read the data and generate the test results and figures.
- We highly recommend using `Python 3.6+` and standard libraries like `numpy, Matplotlib, pandas`. You can choose to use your favourite programming language however the TAs will only be able to assist you with doubts related to Python.
- You are supposed to write your own algorithms, any library functions which implement these directly are strictly off the table. Using them will result in a straight zero on coding questions, `import` wisely!
- **Please start early and clear all doubts ASAP.**
- Please note that the TAs will **only** clarify doubts regarding problem statements. The TAs won't discuss any prospective solution or verify your solution or give hints.
- Post your doubt only on Moodle so everyone is on the same page.

1. [**Regression**] You will implement linear regression as part of this question for the dataset provided. For each sub-question, you are expected to report the following - (i) plot of the best fit curve, (ii) equation of the best fit curve along with coefficients, (iii) value of final least squared error over the test data and (iv) scatter plot of model output vs expected output and for both train and test data. You can also generate a `.csv` file with your predictions on the test data which we should be able to reproduce when we run your command-line instruction.

   Note that you can only regress over the points in the train dataset and you are not supposed to fit a curve on the test dataset. Whatever solution you get for the train data, you have to use that to make predictions on the test data and report results.

(a) (2 marks) Use standard linear regression to get the best fit curve. Vary the maximum degree term of the polynomial to arrive upon an optimal solution.

Solution: (I) Best Fit Curve at Degree 5

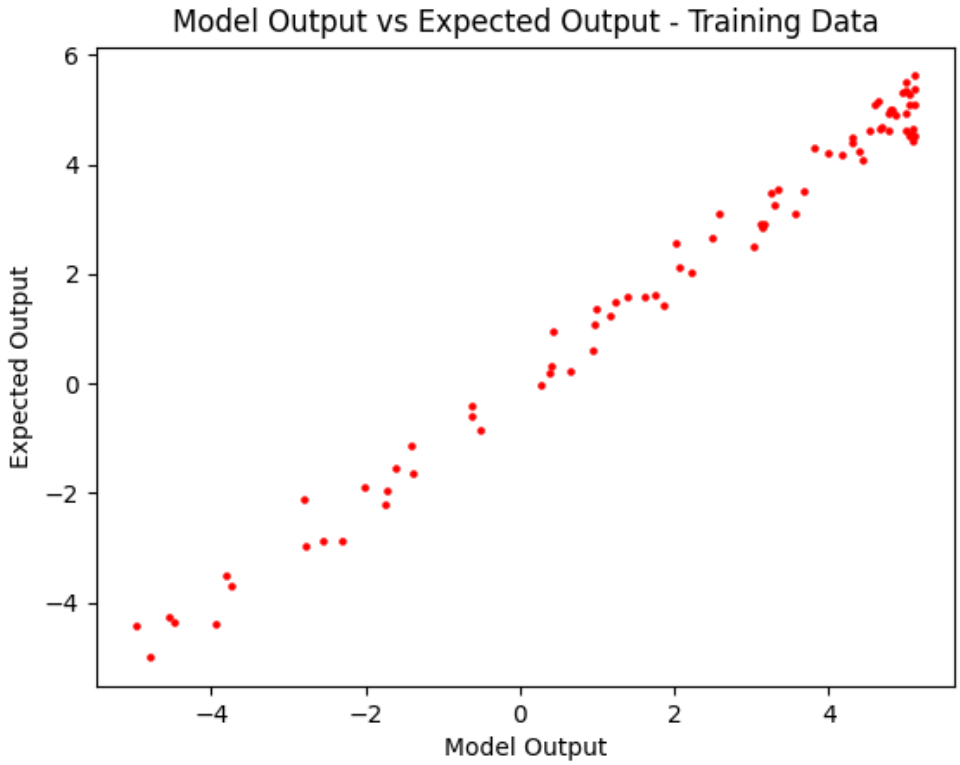


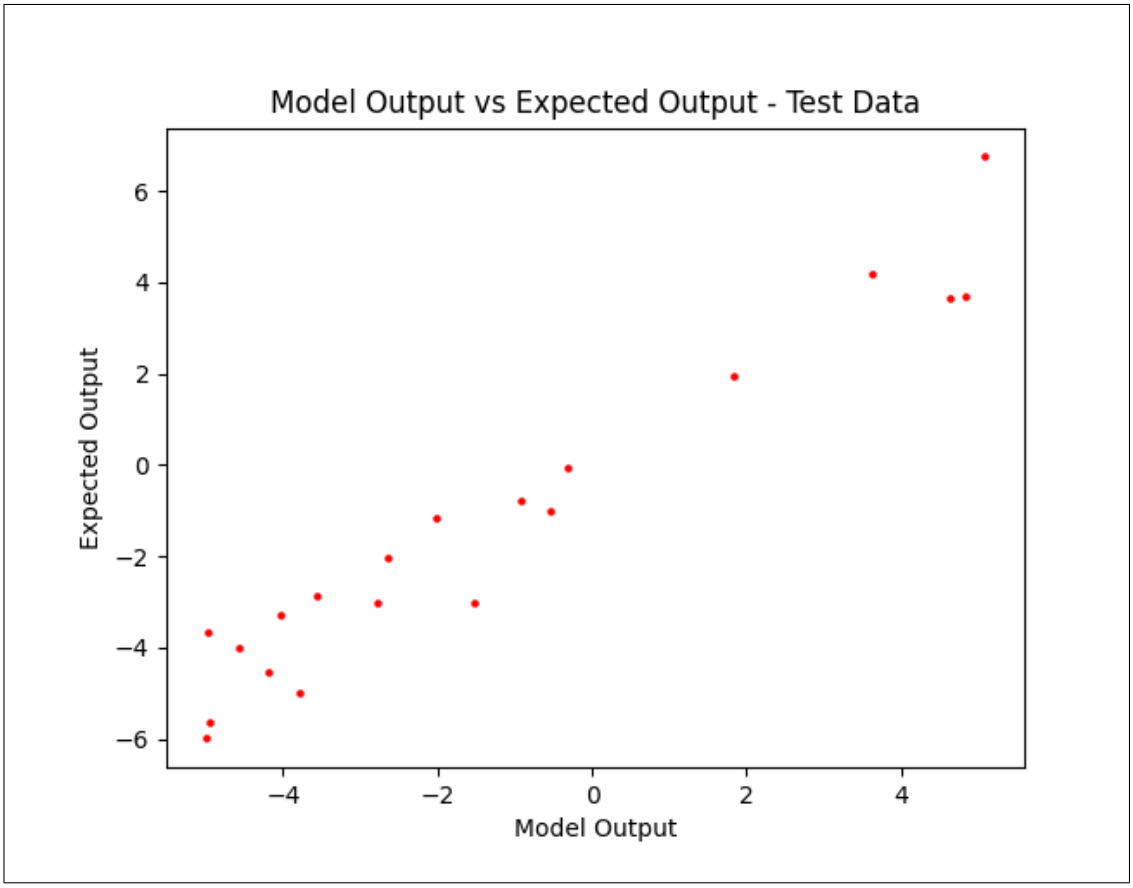
(II) The parameter $w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_0 \end{bmatrix} = \begin{bmatrix} 2.78941392 \\ 5.33713746 \times 10^{-1} \\ -4.92997130 \times 10^{-1} \\ 6.86379156 \times 10^{-2} \\ -2.75075998 \times 10^{-3} \\ 1.21867224 \times 10^{-1} \end{bmatrix}$

$\mathbf{y} = 2.78941392 \;\; x + 5.33713746 \times 10^{-1} x^2 + -4.92997130 \times 10^{-1} x^3$
$+6.86379156 \times 10^{-2} x^4 + -2.75075998 \times 10^{-3} x^5 + 1.21867224 \times 10^{-1}$

(III) Least Squared Error on test data at degree 5 : 0.00026361

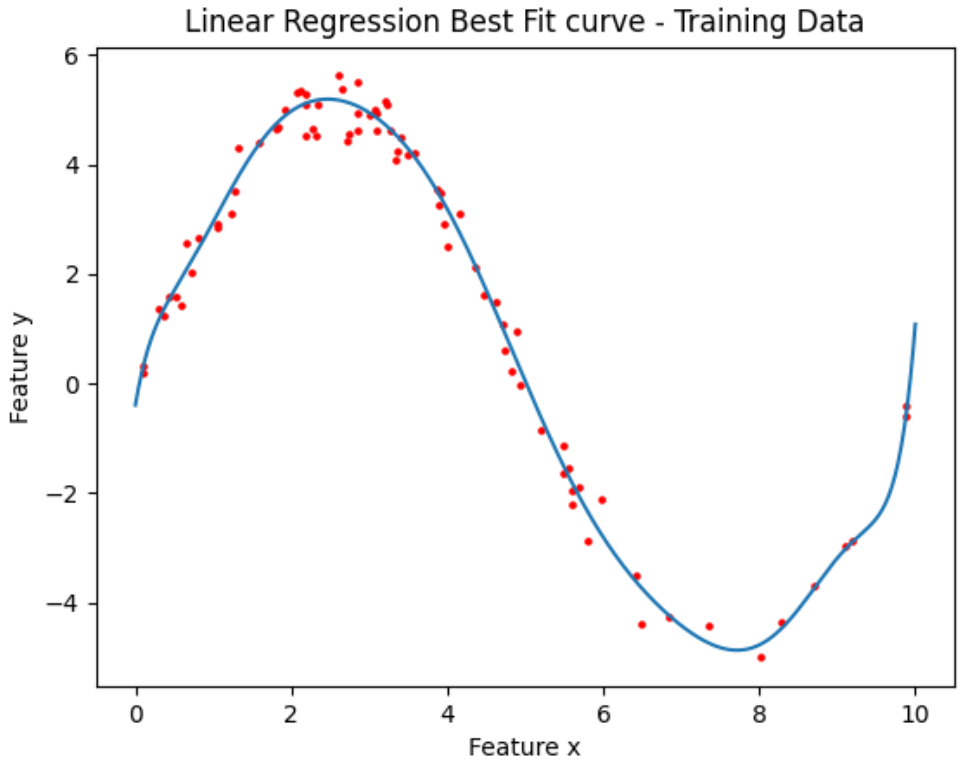
(IV)


Model Output vs Expected Output - Training Data
Expected Output
Model Output
6
4
2
0
−2
−4
−4
−2
0
2
4



(b) (1 mark) In the above problem, increase the maximum degree of the polynomial such that the curve overfits the data.

Solution: (I) We reached Over fitting at degree 11



(II) The parameter $w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ w_7 \\ w_8 \\ w_9 \\ w_10 \\ w_11 \\ w_0 \end{bmatrix} = \begin{bmatrix} -1.61482012 \times 10^1 \\ 2.19295090 \times 10^1 \\ -1.66255571 \times 10^1 \\ 7.57060240 \\ -2.19650555 \\ 4.15148616 \times 10^{-1} \\ -5.08344391 \times 10^{-2} \\ 3.88427159 \times 10^{-3} \\ -1.68211137 \times 10^{-4} \\ 3.15100662 \times 10^{-6} \\ -3.92315045 \times 10^{-1} \\ 8.45327805 \end{bmatrix}$

$\mathbf{y} = -1.61482012 \times 10^1 \ \ x + 2.19295090 \times 10^1 x^2 + -1.66255571 \times 10^1 x^3 +$

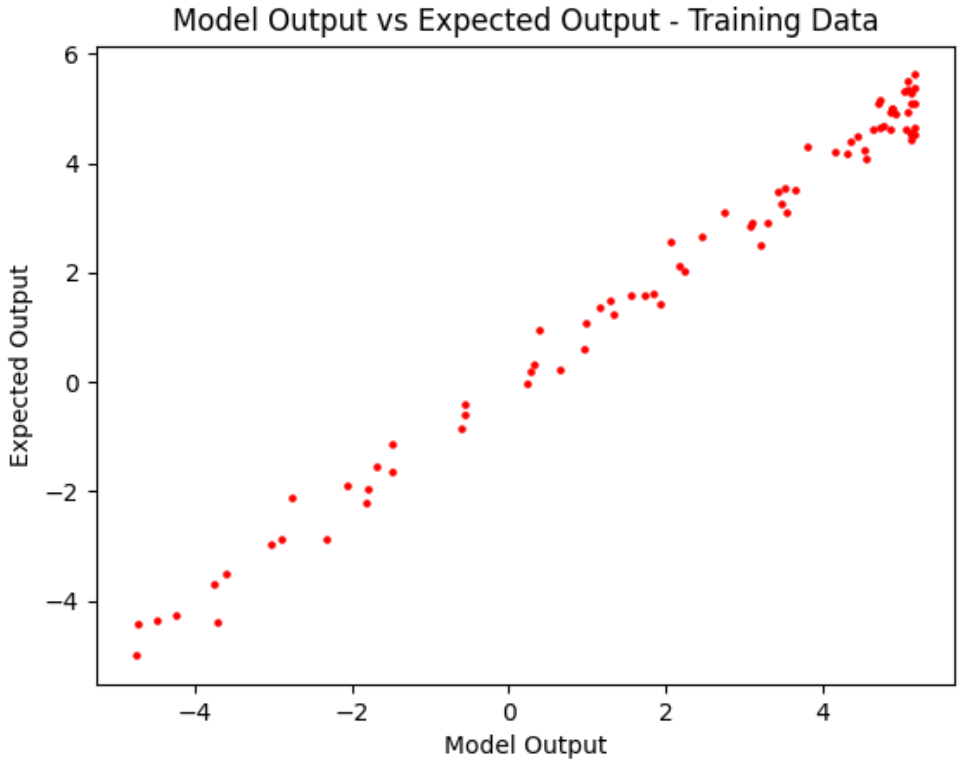
$7.57060240x^4 + -2.19650555x^5 + 4.15148616 \times 10^{-1} x^6 +$

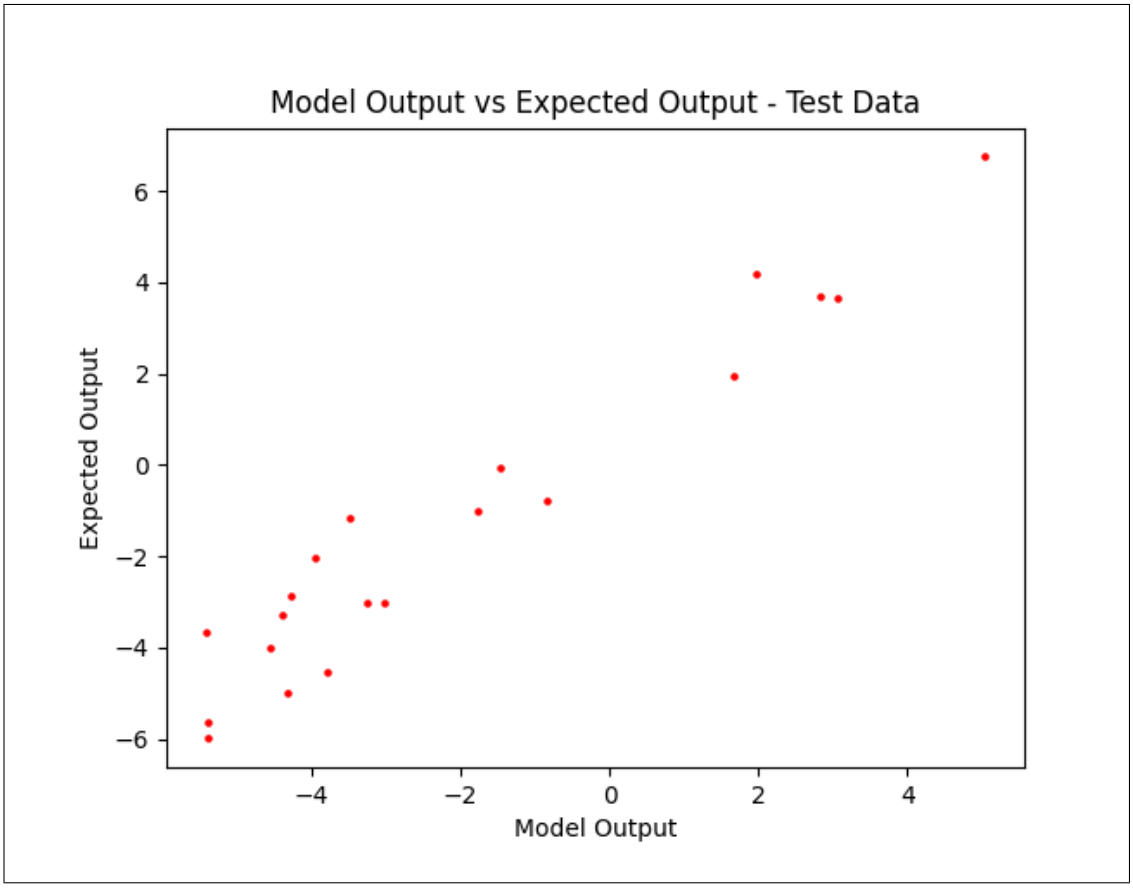
$-5.08344391 \times 10^{-2} x^7 + 3.88427159 \times 10^{-3} x^8 + -1.68211137 \times 10^{-4} x^9 +$

$3.15100662 \times 10^{-6} x^{10} + -3.92315045 \times 10^{-1} x^{11} + 8.45327805$

(III) Least Squared Error on test data at degree 11 : $1.55635754 \times 10^{-5}$

(IV)





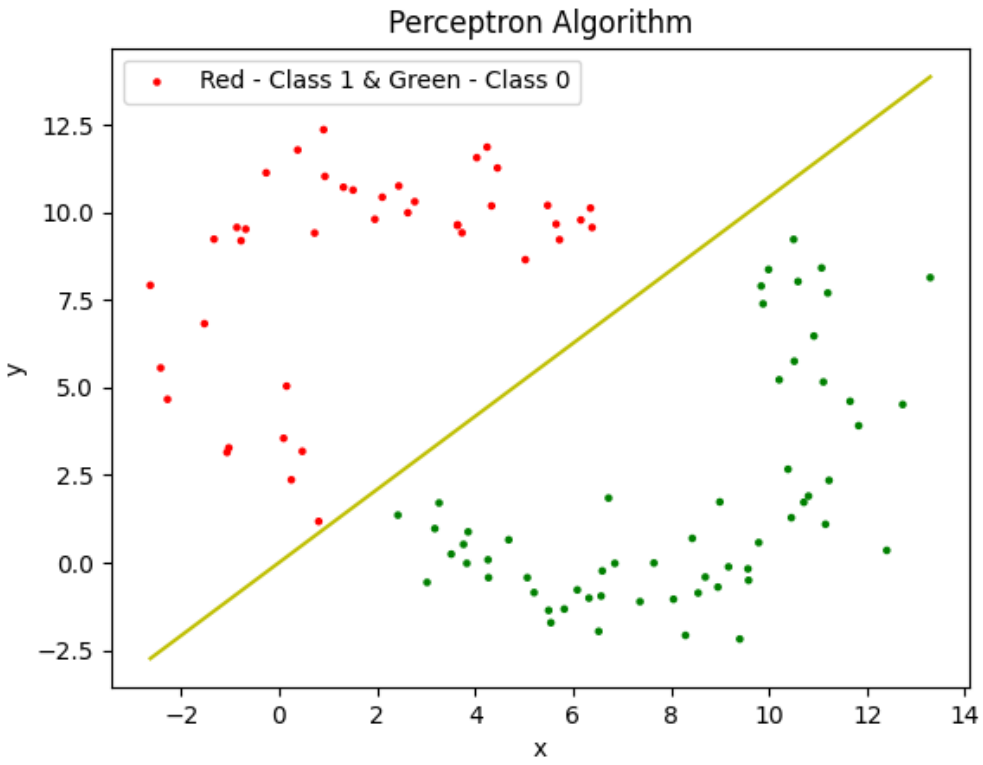
(c) (2 marks) Use ridge regression to reduce the overfit in the previous question, vary the value of lambda (λ) to arrive at the optimal value. Report the optimal λ along with other deliverables previously mentioned.

Solution:

2. **[Classification]** You will implement classification algorithms that you have seen in class as part of this question. You will be provided train and test data as before, of which you are only supposed to use the train data to come up with a classifier which you will use to just make predictions on the test data. For each sub-question below, plot the test data along with your classification boundary and report confusion matrices on both train and test data. Again, your code should generate a `.csv` file with your predictions on the test data as before.

(a) (2 marks) Implement the Preceptron learning algorithm with starting weights as $\mathbf{w} = [0, 0, 1]^T$ for $\mathbf{x} = [1, x, y]^T$ and with a margin of 1.

Solution:

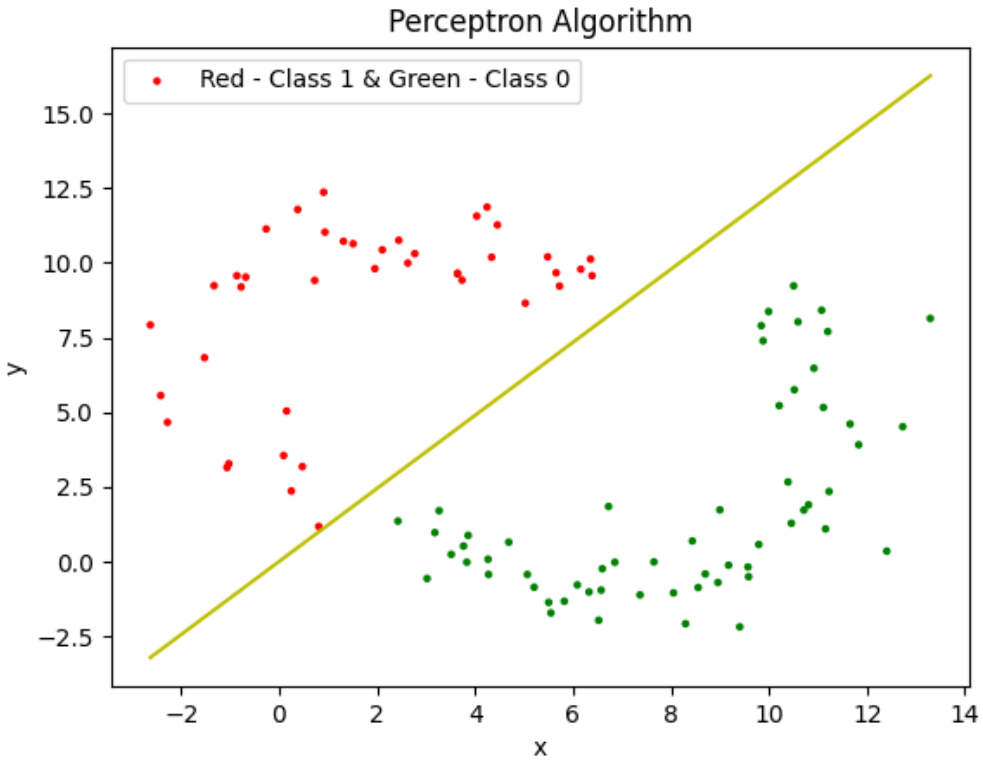


Training data Confusion Matrix $= \begin{bmatrix} 110 & 0 \\ 0 & 90 \end{bmatrix}$

Test data Confusion Matrix $= \begin{bmatrix} 41 & 0 \\ 0 & 59 \end{bmatrix}$

(b) (1 mark) Calculate (code it up!) a Discriminant Function for the two classes assuming Normal distribution when the covariance matrices for both the classes are equal and $C_1 = C_2 = \sigma^2 I$ for some σ.

Solution:

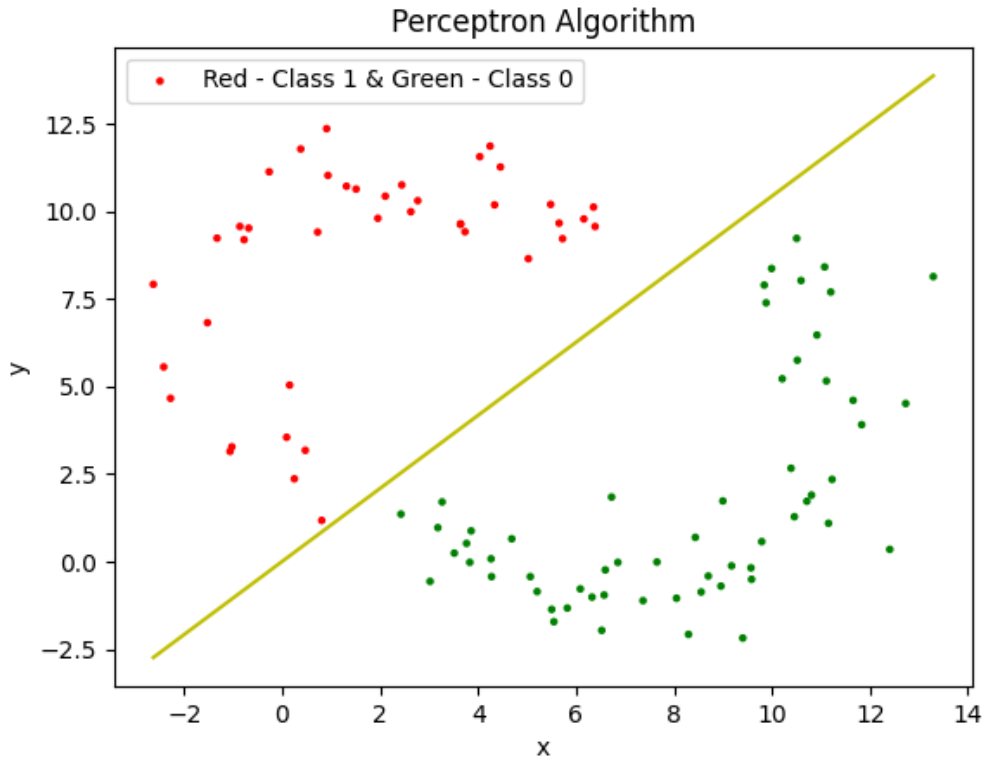


Training data Confusion Matrix $= \begin{bmatrix} 110 & 0 \\ 0 & 90 \end{bmatrix}$

Test data Confusion Matrix $= \begin{bmatrix} 40 & 1 \\ 1 & 58 \end{bmatrix}$

(c) (1 mark) Calculate a Discriminant Function for the two classes assuming Normal distribution when both C_1 and C_2 are full matrices and $C_1 = C_2$.

Solution:



Training data Confusion Matrix $= \begin{bmatrix} 110 & 0 \\ 0 & 90 \end{bmatrix}$

Test data Confusion Matrix $= \begin{bmatrix} 40 & 1 \\ 0 & 59 \end{bmatrix}$

(d) (1 mark) Calculate a Discriminant Function for the two classes assuming Normal distribution when both C_1 and C_2 are full matrices and $C_1 \neq C_2$.

Solution:

3. **[Probability]** In this question, you are required to verify if the following probability mass functions over their respective supports S follow the following properties:

   1. $P(X = x) \geq 0 \quad \forall x \in S$, and
   2. $\sum_{x \in S} P(X = x) = 1$.

   In addition, find the expectation, $\mathbb{E}(X)$ and variance, $Var(X)$ in the following cases.

   (a) (2 marks) A discrete random variable X is said to have a Geometric distribution, with parameter $p \in (0, 1]$ over the support $S = \{1, 2, 3, \dots\}$ if it has the following probability

mass function:

$$P(X = x) = (1 - p)^{x-1}p$$

Solution:
(1) Given $p \in (0, 1]$
$0 < p \leq 1$ ...(1)

$-1 \leq -p < 0$

$0 \leq 1 - p < 1 \quad [\because \ adding \ 1]$

$0 \leq (1 - p)^{x-1} < 1 \quad [\because x > 0]$...(2)

Multiplying (1) and (2),

$0 \leq (1 - p)^{x-1}p < 1$

So $P(X = x) \geq 0 \ \ \forall x \in S$

(2) $\sum_{x \in S} P(X = x) = 1$

$LHS = \sum_{x \in S} P(X = x)$

$= \sum_{x \in S}(1 - p)^{x-1}p$

$= p \sum_{x=1}^{\infty}(1 - p)^{x-1}$

$= p((1 - p)^0 + (1 - p) + ...)$

$= p\frac{1}{1-(1-p)}$

$= p\frac{1}{p}$

$= 1$

$= RHS$

$\mathbf{E[X]} = \sum_{x \in S} xP(X = x)$

$= \sum_{x \in S} x(1 - p)^{x-1}p$

$= p \sum_{x \in S} x(1-p)^{x-1}$

$= p(-\frac{d}{dx} \sum_{x=1}^{\infty} (1-p)^x)$

$= -p(\frac{d}{dx} \frac{1-p}{1-(1-p)})$

$= -p(\frac{d}{dx} \frac{1-p}{p})$

$= p(\frac{d}{dx}(1 - \frac{1}{p}))$

$= pp^{-2}$

$\mathbf{E[X]} = \frac{1}{p}$

$\mathbf{E}[X^2] = \sum_{x \in S} x^2 P(X = x)$

$= \sum_{x \in S} x^2 P(X = x)$

$= \sum_{x \in S} x^2 (1-p)^{x-1} p$

$= p \sum_{x \in S} x((x-1)+1)(1-p)^{x-1}$

$= p \sum_{x \in S} (1-p)^{x-1} x(x-1) + \sum_{x \in S} xp(1-p)^{x-1}$

$= p(1-p) \sum_{x=1}^{\infty} (1-p)^{x-2} x(x-1) + E[X]$

$= p(1-p) \frac{d^2}{dx^2} \sum_{x=1}^{\infty} (1-p)^x + E[X]$

$= p(1-p) \frac{d^2}{dx^2} \frac{1-p}{1-(1-p)} + \frac{1}{p}$

$= p(1-p) \frac{d^2}{dx^2} \frac{1}{p} - 1 + \frac{1}{p}$

$= p(1-p) 2p^{-3} + \frac{1}{p}$

$= \frac{(1-p)2}{p^2} + \frac{p}{p^2}$

$\mathbf{E}[X^2] = \frac{2-p}{p^2}$

$\mathbf{Var[X]} = E[X^2] - [E[X]]^2$

$= \frac{2-p}{p^2} - \frac{1}{p^2}$

Var[X] $= \frac{1-p}{p^2}$

(b) (2 marks) A discrete random variable X is said to have a Poisson distribution, with parameter $\lambda > 0$ over the support $S = \{0, 1, 2, \dots\}$ if it has the following probability mass function:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Solution: (1)

$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$

Term $\frac{\lambda^x}{x!} > 0$ for $\lambda > 0$, $x \geq 0$ and

$e^{-\lambda} \geq 0$ for $\lambda > 0$

So $P(X = x) \geq 0 \quad \forall x \in S$

(2)

$\sum_{x \in S} P(X = x) = 1$

LHS $= \sum_{x \in S} P(X = x)$

$= \sum_{x \in S} \frac{\lambda^x e^{-\lambda}}{x!}$

$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$

$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$

$= \frac{e^{-\lambda}}{e^{-\lambda}}$

$= 1$

$=$ RHS

E[X] $= \sum_{x \in S} xP(X = x)$

$= \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!}$

$= 0 + \sum_{x=1}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!}$ [First term is zero since x = 0]

$= \sum_{y=0}^{\infty} (y+1) \frac{\lambda^{(y+1)} e^{-\lambda}}{(y+1)!}$ [By changing variable: y = x - 1]

$= \sum_{y=0}^{\infty} (y+1) \frac{\lambda \lambda^y e^{-\lambda}}{(y+1)y!}$ [Since (y+1)! = (y+1)y!]

$= \lambda \sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!}$

$\mathbf{E[X]} = \lambda$ [From **(2)** , $\sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!} = 1$]

$\mathbf{E}[X^2] = \sum_{x \in S} x^2 P(X = x)$

$= \sum_{x=0}^{\infty} x^2 \frac{\lambda^x e^{-\lambda}}{x!}$

$= 0 + \sum_{x=1}^{\infty} x^2 \frac{\lambda^x e^{-\lambda}}{x!}$ [First term is zero since x = 0]

$= \sum_{y=0}^{\infty} (y+1)^2 \frac{\lambda^{(y+1)} e^{-\lambda}}{(y+1)!}$ [By changing variable: y = x - 1]

$= \sum_{y=0}^{\infty} (y+1)^2 \frac{\lambda \lambda^y e^{-\lambda}}{(y+1)y!}$ [Since (y+1)! = (y+1)y!]

$= \lambda \sum_{y=0}^{\infty} (y+1) \frac{\lambda^y e^{-\lambda}}{y!}$

$= \lambda \sum_{y=0}^{\infty} (y+1) P(Y = y)$

$= \lambda \sum_{y=0}^{\infty} y P(Y = y) + \lambda \sum_{y=0}^{\infty} P(Y = y)$

$= \lambda(E[Y] + 1)$ [From **(2)**]

$= \lambda(\lambda + 1)$ [From E[X], changing variable X to Y]

$\mathbf{E}[X^2] = \lambda^2 + \lambda$

$\mathbf{Var[X]} = E[X^2] - [E[X]]^2$

$= \lambda^2 + \lambda - (\lambda)^2$

$\mathbf{Var[X]} = \lambda$

4. **[Linear Regression]** Recall the closed form solution for linear regression that we derived in class, the following questions are a follow-up to the same.

   (a) (2 marks) Say we have a dataset where every datapoint has a weight identified with it. Then we have the error function (sum of squares) given by

$$E(w) = \sum_{j=1}^{N} \frac{q_j(y_j - w^T x_j)^2}{2}$$

   where q_j is the weight associated with each of the datapoints ($q_j > 0$). Derive the closed form solution for w^*.

Solution:

$$E(w) = \sum_{j=1}^{N} \frac{q_j(y_i - w^T x_i)^2}{2}$$

here first we will convert above formula to the matrix form.
so here X matrix which will have each row will represent data , matrix Q will be diagonal matrix where each diagonal entry will represent q_i and W will be weight vector and Y is output vector.

$$E(w) = \sum_{j=1}^{N} \frac{q_j(y_i - w^T x_i)^2}{2}$$

$$E(w) = \frac{(Y - XW)^T Q(Y - XW)}{2}$$

$$E(w) = \frac{(Y^T - W^T X^T) Q(Y - XW)}{2}$$

$$E(w) = \frac{(Y^T QY - W^T X^T QY - Y^T QXW + W^T X^T QXW)}{2}$$

Now let's take partial derivative with respect to W and equate to 0 to find solution for w^*.

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

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

$$0 = \frac{\partial(Y^TQY - W^TX^TQY - Y^TQXW + W^TX^TQXW)}{\partial W}$$

$$0 = 0 - X^TQY - X^TQY + 2X^TQXW$$

$$0 = -2X^TQY + 2X^TQXW$$

$$0 = -X^TQY + X^TQXW$$

$$X^TQY = X^TQXW$$

$$W^* = (X^TQX)^{-1}X^TQY$$

(b) (1 mark) We saw in class that the error function in case of ridge regression is given by:

$$\frac{1}{2}\sum_{n=1}^{N}(t_n - w^T\phi(x_n))^2 + \frac{\lambda}{2}w^Tw$$

Show that this error is minimized by :

$$w^* = (\lambda I + \phi^T\phi)^{-1}\phi^Tt$$

Also show that $(\lambda I + \phi^T\phi)$ is invertible for any $\lambda > 0$.

Solution: Now we will convert above equation into the matrix form .
here T is matrix that contain t_i and ϕ will have value $\phi(x_i)$

$$\frac{1}{2}\sum_{n=1}^{N}(t_n - w^T\phi(x_n))^2 + \frac{\lambda}{2}w^Tw = \frac{1}{2}(T - \Phi w)^T(T - \Phi w) + \frac{\lambda}{2}w^Tw$$

$$= \frac{1}{2}((T^T - w^T\Phi^T)(T - \Phi w) + \lambda w^Tw)$$

$$= \frac{1}{2}((T^T - w^T\Phi^T T - T^T\phi w + w^T\Phi^T\Phi w) + \lambda w^T w)$$

Now let's partially derivative by w and equate it to 0 to find w^*

$$\frac{1}{2}\frac{\partial((T^T - w^T\Phi^T T - T^T\phi w + w^T\Phi^T\Phi w + \lambda w^T w)}{\partial w} = 0$$

$$\frac{\partial((T^T - w^T\Phi^T T - T^T\phi w + w^T\Phi^T\Phi w + \lambda w^T w)}{\partial w} = 0$$

$$(-\Phi^T T - \phi^T T + 2\Phi^T\Phi w + 2\lambda w) = 0$$

$$(-2\phi^T T + 2\phi^T\phi w + 2\lambda w) = 0$$

$$2\phi^T\phi w + 2\lambda w = 2\phi^T T$$

$$\phi^T\phi w + \lambda w = \phi^T T$$

$$(\Phi^T\Phi + \lambda I)w = \phi^T T$$

$$w^* = (\Phi^T\phi + \lambda I)^{-1}\phi^T T$$

$$w^* = (\phi^T\phi + \lambda I)^{-1}\phi^T T$$

Hence proved.

Now we need to prove that $(\phi^T\phi + \lambda I)$ is invertible.

Now let's see $\phi^T\phi$ will be positive symmetric matrix and we know positive symmetric matrix are invertible that and $\lambda > 0$ so λI is also positive diagonal matrix so inverse of λI will also positive .
positive inverse means each value in matrix inverse will be positive.

so here $\phi^T\phi$ will have positive inverse and λI will also have positive inverse that means $(\phi^T\phi + \lambda I)$ will also have inverse.

Hence proved, $(\phi^T\phi + \lambda I)$ is invertible.

(c) (1 mark) Given

$$X = \begin{bmatrix} -2 & 6 \\ -1 & 3 \end{bmatrix} \quad y = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

Solve $X^TXw = X^Ty$ such that the Euclidean norm of the solution w^* is minimum.

Solution: here we need to find w.

$$X^TX = \begin{bmatrix} -2 & -1 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} -2 & 6 \\ -1 & 3 \end{bmatrix} = \begin{bmatrix} 5 & -15 \\ -15 & 45 \end{bmatrix}$$

$$X^Ty = \begin{bmatrix} -2 & -1 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \begin{bmatrix} -5 \\ 15 \end{bmatrix}$$

NOW solve $X^TXw = X^Ty$

$$\begin{bmatrix} 5 & -15 \\ -15 & 45 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -5 \\ 15 \end{bmatrix}$$

Now let's do row operation.
$R_2 = R_2 + 3R_1$

$$\begin{bmatrix} 5 & -15 \\ -15 & 45 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -5 \\ 15 \end{bmatrix}$$

$$\begin{bmatrix} 5 & -15 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -5 \\ 0 \end{bmatrix}$$

Now if we back substitute then we will get $5w_1 - 15w_2 = -5$
$5w_1 - 15w_2 = -5$
$w_1 - 3w_2 = -1$
$w_1 = -1 + 3w_2$

so if we take one of the value either w1 or w2 we will get other value according.
so let's take $w_2 = k$ so we will get $w_1 = -1 + 3k$

Now we need to minimize W , so norm(W) $= |W|_2 = \sqrt{w_1^2 + w_2^2} = \sqrt{k^2 + (-1+3k)^2}$ so need to minimize this $|W|_2 = \sqrt{k^2 + (-1+3k)^2}$ so we will do partial derivative with respect to k and equate to 0.

$$\frac{\partial\sqrt{k^2 + (-1+3k)^2}}{\partial k} = 0$$

$$\frac{-3+10k}{\sqrt{k^2+(-1+3k)^2}}=0$$

$$-3+10k=0$$

$$10k=3$$

$$k=\frac{3}{10}$$

here k is at extrem but we don't know that min or max so we need to 2nd derivative to decide min or max.

$$\frac{\partial^2\sqrt{k^2+(-1+3k)^2}}{\partial^2 k}$$

$$\frac{\partial\frac{-3+10k}{\sqrt{k^2+(-1+3k)^2}}}{\partial k}$$

$$\frac{-3+10k}{-3+10k\times(k^2+(-1+3k)^2)^{\frac{2}{3}}}$$

$$\frac{1}{(k^2+(-1+3k)^2)^{\frac{2}{3}}}$$

$$\frac{1}{(k^2 + (-1+3k)^2)^{\frac{2}{3}}} \geq 0$$

so we can see second derivative ≥ 0 so we can tell that it's minimum value for k.
so at $k = \frac{3}{10} = 0.3$ the W is minimum.
$w_2 = k = 0.3$ and $w_1 = 3k - 1 = 3(0.3) - 1 = 0.9 - 1 = -0.1$

$W^* = \begin{bmatrix} -0.1 \\ 0.3 \end{bmatrix}$

5. (2 marks) **[Naive Bayes]** For multiclass classification problems, $p(C_k|\mathbf{x})$ can be written as:

$$p(C_k|\mathbf{x}) = \frac{exp(a_k)}{\sum_j exp(a_j)}$$

where $a_k = \ln p(\mathbf{x}|C_k)p(C_k)$. The above form is called the normalized exponential or softmax function. Now, consider a K class classification problem for which the feature vector $\mathbf{x}$ has M components. Each component is a categorical variable and takes one of L possible values. Let these components be represented using one-hot encoding. Let us also make the naive Bayes assumption that the features are independent given the class. Show that the quantities a_k are linear functions of the components of $\mathbf{x}$.

Solution:

$$p(C_k|\mathbf{x}) = \frac{exp(a_k)}{\sum_j exp(a_j)} = \frac{e^{a_k}}{\sum_j exp(a_j)}$$

here we have given that $a_k = \ln p(\mathbf{x}|C_k)p(C_k)$

$$p(C_k|\mathbf{x}) = \frac{e^{\ln p(\mathbf{x}|C_k)p(C_k)}}{\sum_j exp(a_j)} = \frac{p(\mathbf{x}|C_k)p(C_k)}{\sum_j exp(a_j)}$$

here x is vector which has total M dimension means $x = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_M \end{bmatrix}$

Each component is a categorical variable and takes one of L possible values

and it is represented using one-hot encoding , as we know that in Naive bayes we assume that each feature is independent of thore feature so we can find probability of class given each feature and multiply all that probability to get class given the feature vector.

$$p(C_k|\mathbf{x}) = p(C_k|x_1, x_2, \cdots, x_M) = \frac{p(x_1, x_2, \cdots, x_M|C_k)p(C_k)}{\sum_j exp(a_j)}$$

$$p(C_k|x_1, x_2, \cdots, x_M) = \frac{p(x_1, x_2, \cdots, x_M|C_k)p(C_k)}{\sum_j exp(a_j)}$$

$$p(x_1|C_k)p(x_2|C_k)p(x_3|C_k)\cdots p(x_M|C_k)p(C_k) = \frac{p(x_1, x_2, \cdots, x_M|C_k)p(C_k)}{\sum_j exp(a_j)}$$

here we use assumption that feature are independent.

$$p(x_1|C_k)p(x_2|C_k)p(x_3|C_k)\cdots p(x_M|C_k)p(C_k)\sum_j exp(a_j) = p(x_1, x_2, \cdots, x_M|C_k)p(C_k)$$

now we can all term are in the multiplication we can take log and make it in summation ...

$$\ln p(x_1|C_k)+\ln p(x_2|C_k)+\ln p(x_3|C_k)+\cdots+\ln p(x_M|C_k)p(C_k)+\ln\sum_j exp(a_j) = \ln(p(x_1, x_2, \cdots, x_M|C_k)p(C_k)$$

$$\ln p(x_1|C_k)+\ln p(x_2|C_k)+\ln p(x_3|C_k)+\cdots+\ln p(x_M|C_k)p(C_k)+\ln\sum_j exp(a_j) = \ln(p(\mathbf{x}|C_k)p(C_k))$$

cause $a_k = \ln p(\mathbf{x}|C_k)p(C_k)$ so we put a_k in place of $\ln p(\mathbf{x}|C_k)p(C_k)$

$$\ln p(x_1|C_k) + \ln p(x_2|C_k) + \ln p(x_3|C_k) + \cdots + \ln p(x_M|C_k)p(C_k) + \ln \sum_j exp(a_j) = a_k$$

$$a_k = \ln p(x_1|C_k) + \ln p(x_2|C_k) + \ln p(x_3|C_k) + \cdots + \ln p(x_M|C_k)p(C_k) + \ln \sum_j exp(a_j)$$

Hence we have represented the a_k as the linear functions of the components of $\mathbf{x}$.

Hence proved.

6. (2 marks) **[Naive Bayes]** Consider a Gaussian Naive Bayes classifier for a dataset with single attribute x and two classes 0 and 1. The parameters of the Gaussian distributions are:

$$p(x|y=0) \sim \mathcal{N}(0, 1/4)$$
$$p(x|y=1) \sim \mathcal{N}(0, 1/2)$$
$$P(y=1) = 0.5$$

Find the decision boundary for this classifier if the loss matrix is L = $\begin{bmatrix} 0 & \sqrt{2} \\ 1 & 0 \end{bmatrix}$

Solution: here we need to find the decision boundary classifier.
given loss matrix L = $L = \begin{bmatrix} 0 & \sqrt{2} \\ 1 & 0 \end{bmatrix}$ $P(y=0) = P(y=1)$ are equal.
$P(y=0) = P(y=1) = \frac{1}{2} = 0.5$
so here let's assume that x belog to perticular class and then find the boundary for ot.
let's as x belog to class1.

here for $P(x|y=0)$
$\mu = 0, \sigma^2 = \frac{1}{4}$

$$P(x|y=0) = \frac{1}{\frac{1}{2}\sqrt{2\pi}} e^{-(x-0)^2/2\frac{1}{4}}$$

.

$$P(x|y=0) = \frac{\sqrt{2}}{\sqrt{\pi}} e^{-(x)^2/\frac{1}{2}}$$

.

$R(\alpha_1/x) = L_{11}P(y=1|x) + L_{12}P(y=0|x)$
$R(\alpha_1/x) = 0 \times P(y=1|x) + \sqrt{2}P(y=0|x)$
$R(\alpha_1/x) = \sqrt{2} \times P(y=0|x)$
$R(\alpha_1/x) = \sqrt{2} \times P(x|y=0) \times P(y=0)$
$R(\alpha_1/x) = \sqrt{2} \times P(x|y=0) \times 0.5$
$R(\alpha_1/x) = \sqrt{2} \times \frac{\sqrt{2}}{\sqrt{\pi}} e^{-(x)^2/\frac{1}{2}} \times 0.5$

$R(\alpha_1/x) = \frac{2e^{-2x^2}}{\sqrt{\pi}} \times 0.5$

here for $P(x|y=1)$
$\mu = 0, \sigma^2 = \frac{1}{2}$

$$P(x|y=1) = \frac{1}{\frac{1}{\sqrt{2}}\sqrt{2\pi}} e^{-(x-0)^2/2\frac{1}{2}}$$

.

$$P(x|y=1) = \frac{e^{-x^2}}{\sqrt{\pi}}$$

.

$R(\alpha_0/x) = L_{21}P(y=1|x) + L_{22}P(y=0|x)$
$R(\alpha_0/x) = 1 \times P(y=1|x) + 0 \times P(y=0|x)$
$R(\alpha_0/x) = P(y=1|x)$
$R(\alpha_0/x) = P(x|y=1) \times P(y=0)$
$R(\alpha_0/x) = P(x|y=1) \times 0.5$
$R(\alpha_0/x) = \frac{e^{-x^2}}{\sqrt{\pi}} \times 0.5$

$R(\alpha_0/x) = \frac{e^{-x^2}}{\sqrt{\pi}} \times 0.5$

so here let's assume that x belog to perticular class and then find the boundary for ot. let's as x belog to class1.

$$R(\alpha_0/x) \geq R(\alpha_1/x)$$

$$\frac{e^{-x^2}}{\sqrt{\pi}} \times 0.5 \geq \frac{2e^{-2x^2}}{\sqrt{\pi}} \times 0.5$$

$$\frac{e^{-x^2}}{\sqrt{\pi}} \geq \frac{2e^{-2x^2}}{\sqrt{\pi}}$$

$$e^{-x^2} \geq 2e^{-2x^2}$$

let's apply log on both side

$$\log_e e^{-x^2} \geq \log_e 2 + \log_e e^{-2x^2}$$

$$-x^2 \geq 0.6931 + -2x^2$$

$$x^2 \geq 0.6931$$

$$x \geq 0.8325$$

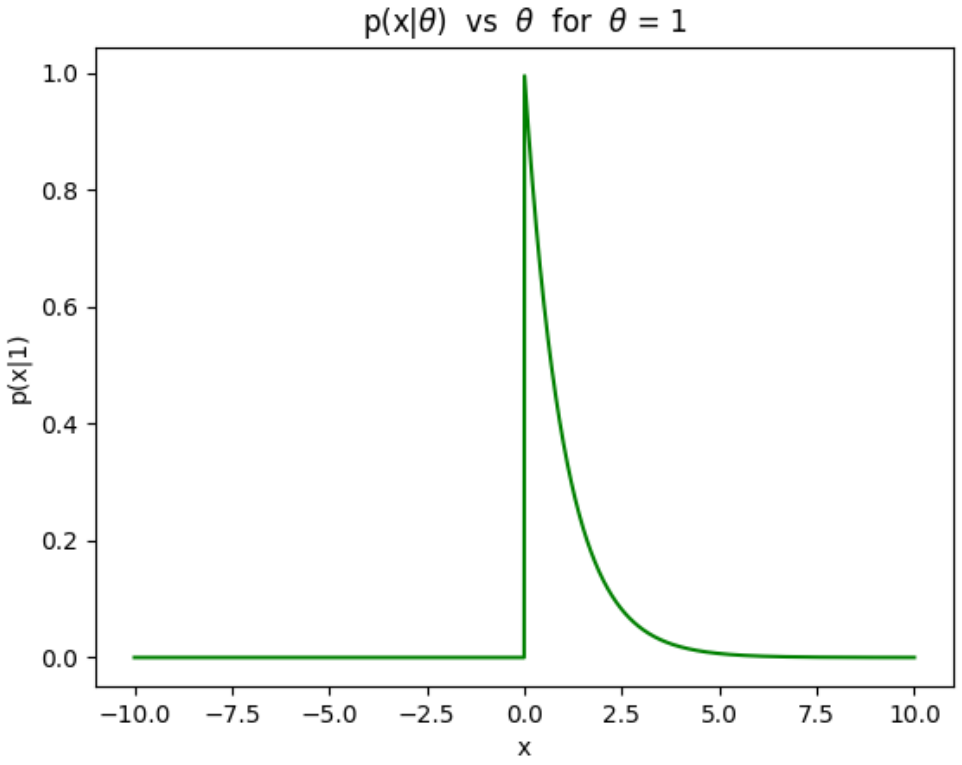
so if $x \geq 0.8325$ then x belong to class 1 else class 0
our decision boundary is 0.8325.

7. **[MLE]** Let x have an exponential density

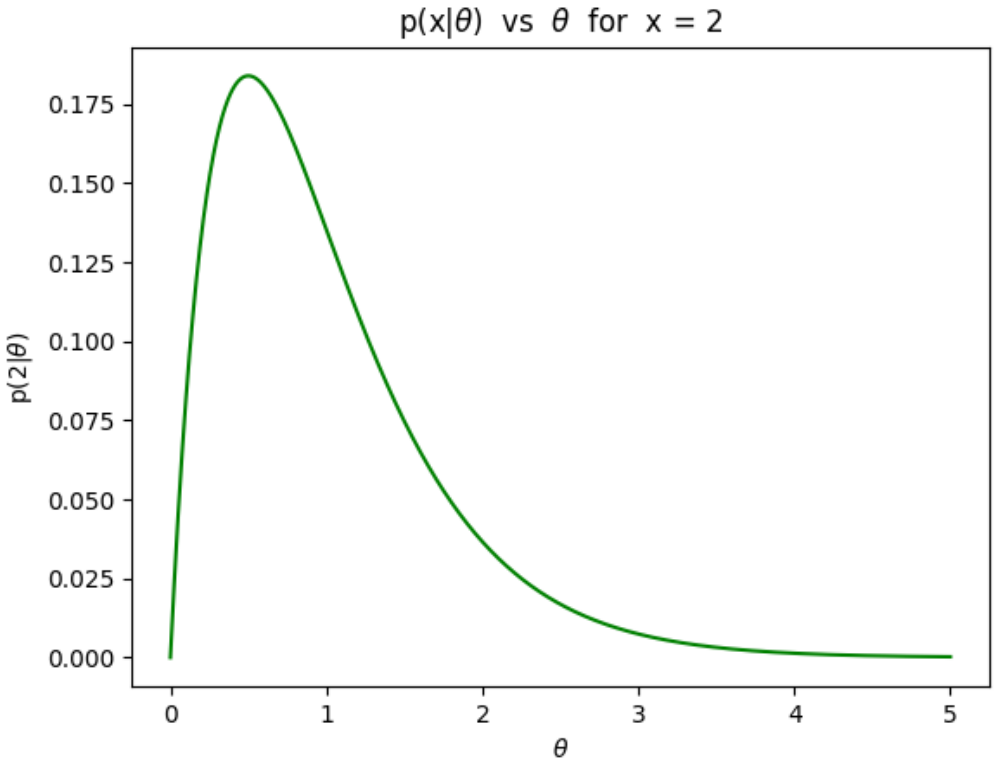
$$p(x|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(a) (2 marks) Plot $p(x|\theta)$ versus x for $\theta = 1$. Plot $p(x|\theta)$ versus θ, $(0 \leq \theta \leq 5)$, for $x = 2$.

Solution: (1)



(2)



(b) (1 mark) Suppose that n samples $x_1, ..., x_n$ are drawn independently according to $p(x|\theta)$. Give the maximum likelihood estimate for θ.

Solution: here $x_1, ..., x_n$ are drawn independently so maximum likelihood will be multiplication of individual sample likelihood, but then we will take the log of that for log-likelihood so all multiplication will convert into the summation.
Then after apply log we will find derivative of that and equate it to 0 , and find θ. Here we can apply log cause derivative of function and derivative of log function both will be 0 at same place,so purpose of finding derivative and equate to 0 we can find best suitable value for θ.
so log likelihood will be $\cdots$

$$L(\theta|x_1, x_2 \cdots x_n) = \log_e(\prod_{i=1}^{n} L(\theta|x_i))$$

$$L(\theta|x_1, x_2 \cdots x_n) = (\sum_{i=1}^{n} \log_e L(\theta|x_i))$$

$$L(\theta|x_1, x_2 \cdots x_n) = (\sum_{i=1}^{n} \log_e \theta e^{-\theta x_i})$$

$$L(\theta|x_1, x_2 \cdots x_n) = (\sum_{i=1}^{n} (\log_e \theta + \log_e e^{-\theta x_i}))$$

$$L(\theta|x_1, x_2 \cdots x_n) = (\sum_{i=1}^{n} (\log_e \theta - \theta x_i))$$

$$L(\theta|x_1, x_2 \cdots x_n) = (n \log_e \theta - \sum_{i=1}^{n} \theta x_i)$$

$$L(\theta|x_1, x_2 \cdots x_n) = (n \log_e \theta - \theta \sum_{i=1}^{n} x_i)$$

$$L(\theta|x_1, x_2 \cdots x_n) = (n \log_e \theta - \theta \sum_{i=1}^{n} x_i)$$

Now let's find the partial derivative of that with respect to θ and equate the derivative to 0.

$$0 = (\frac{n}{\theta} - \sum_{i=1}^{n} x_i)$$

$$\frac{n}{\theta} = \sum_{i=1}^{n} x_i$$

$$\theta = \frac{n}{\sum_{i=1}^{n} x_i}$$

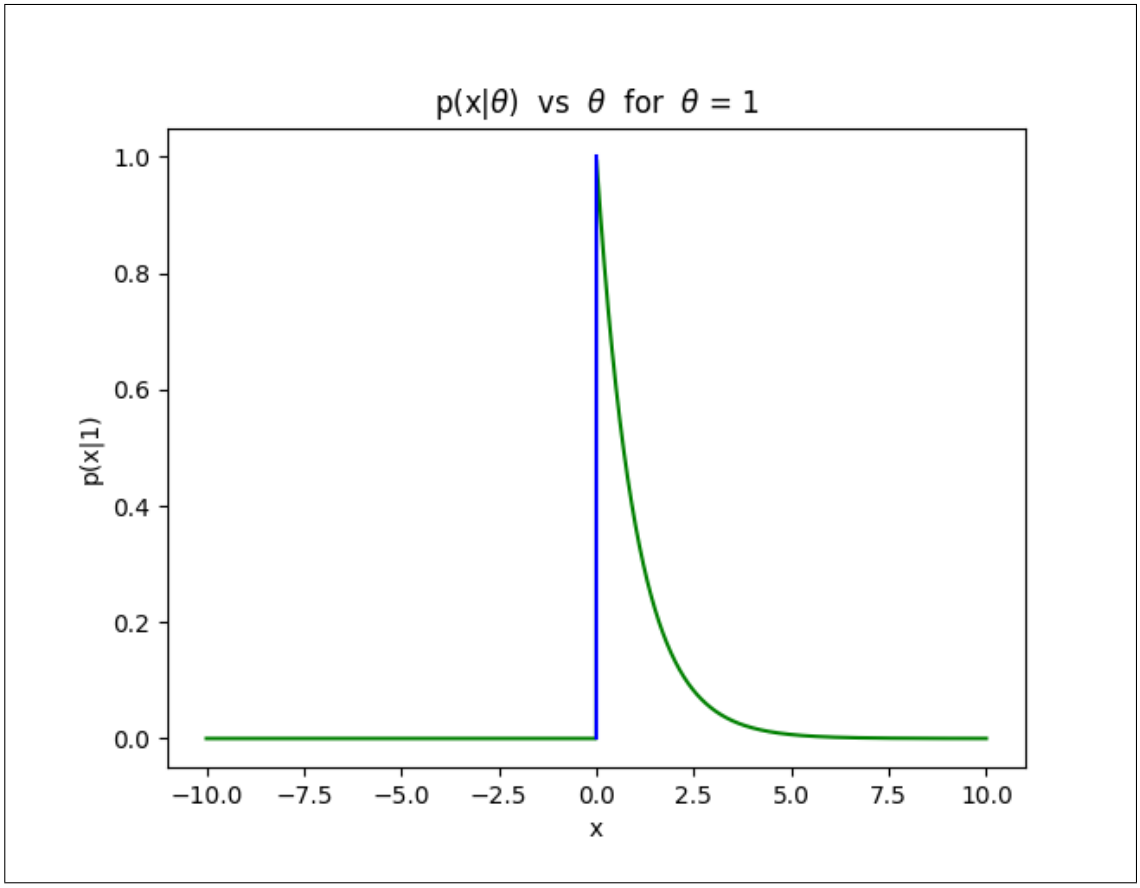
$$\theta = \frac{n}{\sum_{i=1}^{n} x_i}$$

so maximum likelihood estimate for θ is $\cdots$

$$\theta = \frac{n}{\sum_{i=1}^{n} x_i}$$

(c) (2 marks) On the graph generated with $\theta = 1$ in part (a), mark the maximum likelihood estimate $\hat{\theta}$ for large n. Write down your observations.

Solution:



8. (3 marks) **[MLE]** Gamma distribution has a density function as follows

$$f(x|\alpha,\lambda) = \frac{1}{\Gamma(\alpha)}\lambda^{\alpha}x^{\alpha-1}e^{-\lambda x}, \quad \text{with} \quad 0 \leq x \leq \infty$$

Suppose the parameter α is known, please find the MLE of λ based on an i.i.d. sample $X_1, ..., X_n$.

Solution:

$f(x|\alpha,\lambda) = \frac{1}{\Gamma(\alpha)}\lambda^{\alpha}x^{\alpha-1}e^{-\lambda x}, \quad \text{with} \quad 0 \leq x \leq \infty$

As given α is known so need to find λ here.

$$L(x_i) = \frac{1}{\Gamma(\alpha)}\lambda^{\alpha}x_i^{\alpha-1}e^{-\lambda x_i}$$

Now take log on both the side

$$\log_e L(x_i) = -\log_e \Gamma(\alpha) + \alpha log_e \lambda + (\alpha - 1)\log_e x_i - \lambda x_i$$

so Log likelihood L(X) will be summation of all Sample , then we will find partial derivative and equate to zero to find max value suitable valse of λ.

$$L(X) = \sum_{i=1}^{n} \log_e L(x_i)$$

$$L(X) = \sum_{i=1}^{n} (-\log_e \Gamma(\alpha) + \alpha log_e \lambda + (\alpha - 1)\log_e x_i - \lambda x_i)$$

Now partially derivate with respect to λ and equate to 0 to find λ.

$$\frac{\partial L(X)}{\partial \lambda} = \sum_{i=1}^{n} \frac{\partial(-\log_e \Gamma(\alpha) + \alpha log_e \lambda + (\alpha - 1)\log_e x_i - \lambda x_i)}{\partial \lambda}$$

Here $-\log_e \Gamma(\alpha)$ and $(\alpha - 1)\log_e x_i$ term will be 0 cause it did not have any λ term.

$$0 = \sum_{i=1}^{n} (0 + \frac{\alpha}{\lambda} + 0 - x_i)$$

$$0 = \sum_{i=1}^{n}(\frac{\alpha}{\lambda} - x_i)$$

$$0 = n\frac{\alpha}{\lambda} - \sum_{i=1}^{n}(x_i)$$

$$n\frac{\alpha}{\lambda} = \sum_{i=1}^{n}(x_i)$$

$$\lambda = \frac{n\alpha}{\sum_{i=1}^{n}(x_i)}$$

$$\lambda = \frac{\alpha}{\frac{\sum_{i=1}^{n}(x_i)}{n}}$$

$$\lambda = \frac{\alpha}{\overline{X}}$$

Where $\overline{X}$ is mean of all sample data means $\overline{X} = \frac{\sum_{i=1}^{n}(x_i)}{n}$

So MLE of λ will be

$$\lambda = \frac{\alpha}{\overline{X}}$$