CS5691: Pattern Recognition and Machine Learning
Assignment #1

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

Teammate 1: Varun Gumma **Roll number:** CS21M070
Teammate 2: Uday Sai Vemula **Roll number:** CS21M066

- 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.

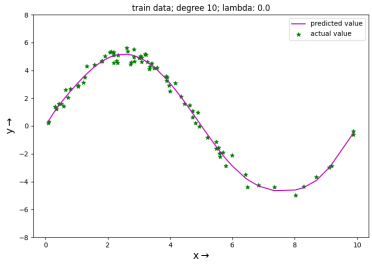
1. [**Regression**] You will implement linear regression as part of this question for the dataset provided. For each sub-question, you are expected to report the following - (i) plot of the best fit curve, (ii) equation of the best fit curve along with coefficients, (iii) value of final least squared 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:

(a) *Plot of best fit:*

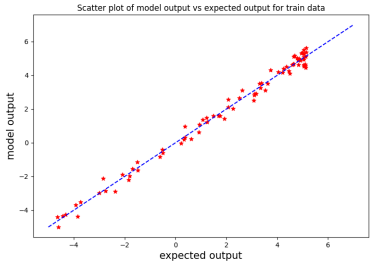


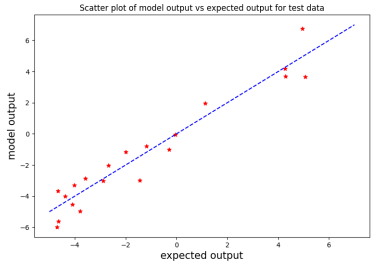
(b) *Equation of best fit:*

$$\texttt{-0.05730404797941446 + 4.514885067939758}x \texttt{ - 3.4465465545654297}x^2\texttt{, + 3.608590930700302}x^3 \texttt{ - 2.3021207749843597}x^4 \texttt{ + 0.8448312655091286}x^5 \texttt{ - 0.19405682757496834}x^6 \texttt{ + 0.028354724403470755}x^7 \texttt{ - 0.0025439037126488984}x^8 \texttt{ + 0.00012715959928755183}x^9 \texttt{ - 0.0000027030937985728087}x^{10}$$

(c) *Final least squared error over test data:* `16.781612363825797`

(d) *Scatter plot of train and test data:*

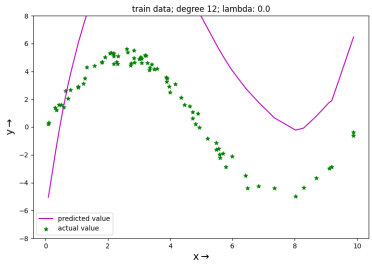




(b) (1 mark) In the above problem, increase the maximum degree of the polynomial such that the curve overfits the data.

Solution:

(a) *Plot of best fit:*

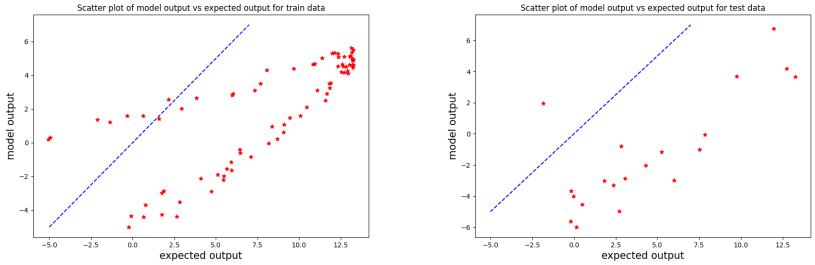


(b) *Equation of best fit:*

$$\texttt{-6.358128573745489 + 13.315819919109344}x \texttt{ + 7.829554438591003}x^2 \texttt{ -23.16979217529297}x^3 \texttt{ + 24.361974954605103}x^4 \texttt{ - 14.927240371704102}x^5 \texttt{ + 5.798273146152496}x^6 \texttt{ - 1.4829664006829262}x^7 \texttt{ + 0.25292173214256763}x^8 \texttt{ - 0.028460026485845447}x^9 \texttt{ + 0.0020274561102269217}x^{10} \texttt{ - 0.00008280010433736607}x^{11} \texttt{ + 0.0000014760687605530 62}x^{12}$$

(c) *Final least squared error over test data:* `823.7308485805207`

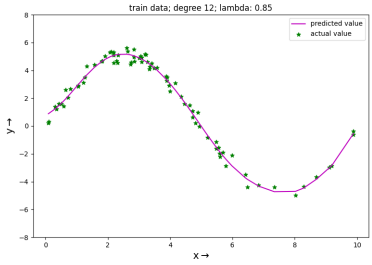
(d) *Scatter plot of train and test data:*



(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:

(a) *Plot of best fit:*

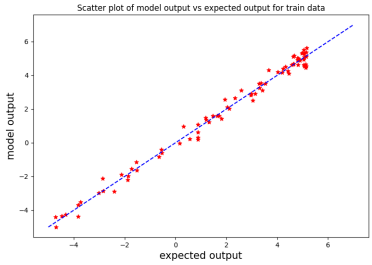


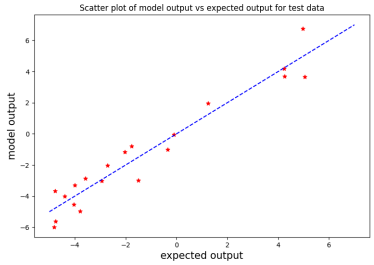
(b) *Equation of best fit:*

$$\texttt{0.7566766768140951 + 1.220936962054111}x \texttt{ + 0.8924203288042918}x^2 \texttt{ + 0.3121445416472852}x^3 \texttt{ - 0.25208010073401965}x^4 \texttt{ - 0.27710046782158315}x^5 \texttt{ + 0.2735783620737493}x^6 \texttt{ - 0.10591288091382012}x^7 \texttt{ + 0.02311308118805755}x^8 \texttt{ - 0.0030654988058813615}x^9 \texttt{ + 0.00024537968579352 31}x^{10} \texttt{ - 0.000010919772542195005}x^{11} \texttt{ + 0.0000020773860254230 67}x^{12}$$

(c) *Final least squared error over test data:* `17.00096739607598`

(d) *Scatter plot of train and test data:*





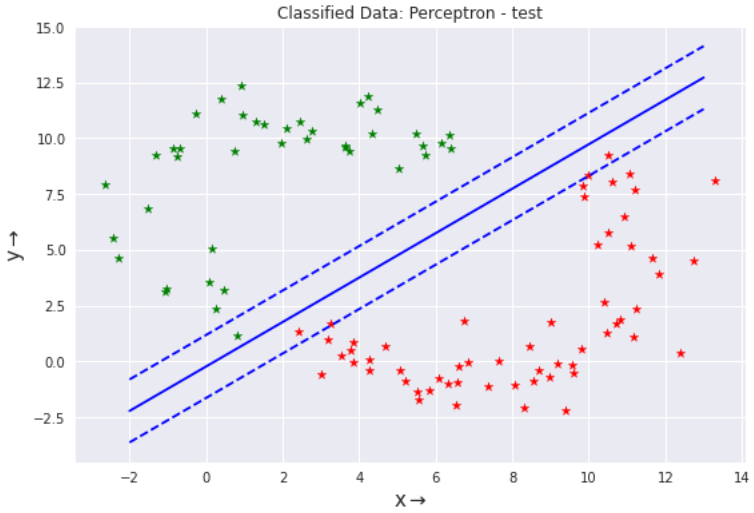
(e) *Optimal λ value:* `0.85`

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 Perceptron 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:

(a) Classification boundary and test data:



(b) Confusion matrices:

```
confusion matrix for train data:
+-------------+
|  90  |  0   |
+-------------+
|  0   | 110  |
+-------------+

confusion matrix for test data:
+-------------+
|  59  |  0   |
+-------------+
|  0   |  41  |
+-------------+
```

(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: Here σ was taken as 2.

(a) Classification boundary and test data:

(b) Confusion matrices:

```
confusion matrix for train data:
+-------------+
|  90  |  0   |
+-------------+
|  0   | 110  |
+-------------+

confusion matrix for test data:
+-------------+
|  58  |  1   |
+-------------+
|  1   |  40  |
+-------------+
```

(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: Here $C_1 = C_2$ both were taken as C_1 (covariance of positive samples)

(a) Classification boundary and test data:

(b) Confusion matrices:

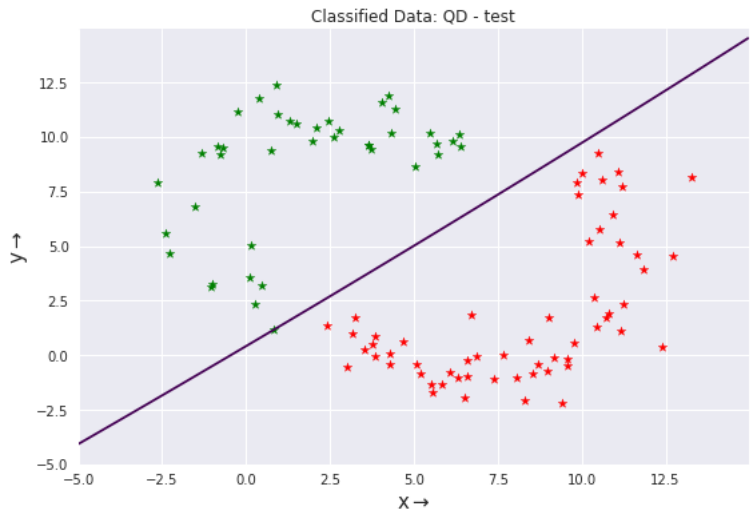
```
confusion matrix for train data:
+-------------+
|  90  |  0   |
+-------------+
|  0   | 110  |
+-------------+

confusion matrix for test data:
+-------------+
|  59  |  0   |
+-------------+
|  0   |  41  |
+-------------+
```

(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:

(a) Classification boundary and test data:



(b) Confusion matrices:

```
confusion matrix for train data:
+-------------+
|  90  |  0   |
+-------------+
|  0   | 110  |
+-------------+

confusion matrix for test data:
+-------------+
|  59  |  0   |
+-------------+
|  0   |  41  |
+-------------+
```

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. $P(X = x) \geq 0$: As $P(X = x) = (1-p)^{x-1}p$. This is product of 2 positive terms as p (probability) > 0 and $(1-p)^{x-1}$ is a exponential which is non-negative (this value can be zero, when $p = 1$). Hence, the product is also non-negative or ≥ 0.

2. $\sum_{x \in S} P(X = x) = 1$: $\sum_{x \in S} P(X = 1) = \sum_{i=1}^{\infty} P(X = i) = \sum_{i=1}^{\infty} p(1-p)^{i-1} = p\sum_{i=0}^{\infty} p(1-p)^{i}$. The summation represents a geometric progression whose sum is $\frac{1}{1-(1-p)} = \frac{1}{p}$. $\therefore$ the total value is $p \cdot \frac{1}{p} = 1$. Hence, proved.

3. $E(X)$: $E(X) = \sum_{x \in S} xP(X = x) = \sum_{i=1}^{\infty} ip(1-p)^{i-1} = p\sum_{i=1}^{\infty} i(1-p)^{i-1} = -p\sum_{i=1}^{\infty} \frac{d}{dp}(1-p)^{i} = -p\sum_{i=0}^{\infty} \frac{d}{dp}(1-p)^{i}$ (here the summation can be extended to 0, as that term is a constant 1 and will become 0 when differentiated and does not contribute to the sum). $\therefore -p\sum_{i=0}^{\infty} \frac{d}{dp}(1-p)^{i} = -p\frac{d}{dp}(\sum_{i=0}^{\infty}(1-p)^{i}) = -p\frac{d}{dp}(\frac{1}{p}) = -p \cdot \frac{-1}{p^2} = \frac{1}{p}$.

4. $Var(X)$: $E(X^2) = \sum_{x \in S} x^2P(X = x) = \sum_{i=1} i^2P(X = i) = \sum_{i=1}^{\infty} i^2p(1-p)^{i-1}$. Here, we can once again extend the summation to 0, as that term will be 0 and will not contribute to the summation even when included.$\therefore \sum_{i=0}^{\infty} i^2p(1-p)^{i-1} = \sum_{i=0}^{\infty}(i^2-i+i)p(1-p)^{i-1} = p(1-p)\sum_{i=0}^{\infty} i(i-1)(1-p)^{i-2} + p\sum_{i=0}^{\infty} i(1-p)^{i-1} = p(1-p)\frac{d^2}{dp^2}\sum_{i=0}^{\infty}(1-p)^{i} + p\frac{d}{dp}\sum_{i=0}^{\infty}(1-p)^{i} = p(1-p) \cdot \frac{2}{p^3} + p \cdot \frac{1}{p^2} = \frac{2-2p+p}{p^2} = \frac{2-p}{p^2}$. $Var(X) = E(X^2) - (E(X))^2 = \frac{2-p-1}{p^2} = \frac{1-p}{p^2}$

(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. $P(X = x) \geq 0$: As $P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$. All terms in this expression are non-negative, as λ^x, $e^{-\lambda}$ are exponentials and positive and $x! \geq 0$ ($\forall x \geq 0$). Hence this PMF ≥ 0.

2. $\sum_{x \in S} P(X = x) = 1$: $\sum_{x \in S} P(X = x) = \sum_{i=0}^{\infty} P(X = i) = \sum_{i=0}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} = e^{-\lambda}\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} \cdot e^{\lambda} = 1$. Hence, proved.

3. $E(X)$: $E(X) = \sum_{x \in S} xP(X = x) = \sum_{i=0}^{\infty} i \cdot \frac{\lambda^i e^{-\lambda}}{i!} = \sum_{i=1}^{\infty} i \cdot \frac{\lambda^i e^{-\lambda}}{i!}$ (with $i = 0$, the first term is 0 and can be neglected) $\therefore \sum_{i=1}^{\infty} i \cdot \frac{\lambda^i e^{-\lambda}}{i!} = \lambda e^{-\lambda}\sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda$.

4. $Var(X)$: $E(X^2) = \sum_{x \in S} x^2P(X = x) = \sum_{i=0}^{\infty} i^2 \cdot \frac{\lambda^i e^{-\lambda}}{i!} = \sum_{i=1}^{\infty} i^2 \cdot \frac{\lambda^i e^{-\lambda}}{i!}$ (with $i = 0$, the first term is 0 and can be neglected). $\therefore \sum_{i=1}^{\infty} i^2 \cdot \frac{\lambda^i e^{-\lambda}}{i!} =$

$\sum_{i=1}^{\infty}(i^2-i+i)\cdot\frac{\lambda^i e^{-\lambda}}{i!} = \sum_{i=1}^{\infty} i(i-1)\cdot\frac{\lambda^i e^{-\lambda}}{i!} + \sum_{i=1}^{\infty} i\cdot\frac{\lambda^i e^{-\lambda}}{i!} = \lambda^2 e^{-\lambda}\sum_{i=2}^{\infty}\frac{\lambda^{i-2}}{(i-2)!} + \lambda e^{-\lambda}\sum_{i=1}^{\infty}\frac{\lambda^{i-1}}{(i-1)!} = (\lambda^2+\lambda)e^{-\lambda}\cdot e^{\lambda} = \lambda^2+\lambda$. $\therefore Var(X) = E(X^2) - (E(X))^2 = \lambda^2+\lambda-\lambda^2 = \lambda$. $\therefore$ for this distribution, mean is same as variance.

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: The error function for weighted-least squares can be re-written as $\frac{1}{2}\sum_{j=1}^{N}(y_j - w^T x_j)q_j(y_j - w^T x_j)$. With X as data matrix, Y as target vector, W as weight vector and Q as a diagonal matrix with the ii^{th} entry being the weight for the i^{th} data point, the function can be vectorized as $E(w) = \frac{1}{2}(Y-Xw)Q(Y-Xw)$.

   Taking a derivative, w.r.t W, the gradient for this error function (vectorized) will be of the form $\nabla E(w) = -\frac{1}{2}\cdot 2\cdot X^T Q(Y-Xw) = -X^T Q(Y-Xw)$. At optimial value of w, the error will be minimum or $\nabla E(w^*) = 0$. $\therefore -X^T Q(Y-Xw^*) = 0$ or $X^T QY = X^T QXw^*$. Left multipling by $(X^T QX)^{-1}$, we have, $w^* = (X^T QX)^{-1}X^T QY$.

   We can observe that this is a generalized version of the closed form solution. If $Q = I$ (each data point has unit weight), we have $w^* = (X^T IX)^{-1}X^T IY = (X^T X)^{-1}X^T Y$, i.e. the solution for ordinary least squares.

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

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

   Show that this error is minimized by :

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

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

   Solution: With ϕ representing the data matrix (with matured features), Y as target vector and W as weight vector, the error function for least squares can be vectorized as

$\frac{1}{2}(t-\phi w)^T(t-\phi w)+\frac{\lambda}{2}w^Tw$. The gradient of this error function w.r.t. w (vectorized) will be $\frac{1}{2}\cdot 2\cdot -\phi^T(t-\phi w)+\frac{\lambda}{2}\cdot 2\cdot w=\phi^T(t-\phi w)+\lambda w$. At the optimum value of w, i.e. w^*, the error will be minimum and gradient of the error function will be zero. $\therefore -\phi^T(t-\phi w^*)+\lambda w^*=0$. $\therefore -\phi^Tt+\phi^T\phi w^*+\lambda w^*=-\phi^Tt+(\phi^T\phi+\lambda I)w^*=0$ or $(\phi^T\phi+\lambda I)w^*=\phi^Tt$ or $w^*=(\phi^T\phi+\lambda I)^{-1}\phi^Tt$.

$\phi^T\phi$ may or may not be invertible. If not, it has some dependent columns and the lacks some pivots. By adding λI to $\phi^T\phi$, we have added a positive value to all the diagonal elements of $\phi^T\phi$, which makes sure there are no zero pivots, i.e. all n-pivots are available and the matrix is full rank and 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: $X^TX=\begin{bmatrix}-2 & -1\\ 6 & 3\end{bmatrix}\begin{bmatrix}-2 & 6\\ -1 & 3\end{bmatrix}=\begin{bmatrix}5 & -15\\ -15 & 45\end{bmatrix}$ and $X^TY=\begin{bmatrix}5 & -15\\ -15 & 45\end{bmatrix}\begin{bmatrix}3\\ -1\end{bmatrix}=\begin{bmatrix}-5\\ 15\end{bmatrix}$. $\therefore$ the w for this system is obtained by solving $\begin{bmatrix}-2 & -1\\ 6 & 3\end{bmatrix}w=\begin{bmatrix}-5\\ 15\end{bmatrix}$ or $5w_1-15w_2=-5$, $-15w_1+45w_2=15$. Since these two represent the same line, we have w as the solution to $w_1-3w_2=-1$ (which are infinite in number). To find, w^*, which is at minimum Euclidean distance (L_2 norm) from the origin, we drop a perpendicular from origin onto $w_1-3w_2=-1$ (this line will become tangent to the circle with radius $||w^*||_2$, which will be $3w_1+w_2=0$. $\therefore$ the point of intersection, i.e. $w^*=\begin{bmatrix}-0.1\\ 0.3\end{bmatrix}$, and it has a norm of $\frac{1}{\sqrt{10}}$.

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: Since given the class label the features are independent, we have $p(\mathbf{x}|C_k)=p(x_1|C_k)p(x_2|C_k)\dots p(x_M|C_k)=\prod_{m=1}^{M}p(x_m|C_k)$. As x_m is a one-hot vector of dimension

L (multinoulli distribution), we $p(x_m|\mu_{km}, C_k) = \prod_{l=1}^{L} \mu_{kml}^{x_{ml}}$ or $p(\mathbf{x}|C_k) = \prod_{m=1}^{M}\prod_{l=1}^{L} \mu_{kml}^{x_{ml}}$

$a_k = \ln p(\mathbf{x}|C_k)p(C_k) = \ln p(\mathbf{x}|C_k) + \ln p(C_k) = \ln \prod_{m=1}^{M}\prod_{l=1}^{L} \mu_{kml}^{x_{ml}} + \ln p(C_k) = \sum_{m=1}^{M}\sum_{l=1}^{L} x_{ml} \ln \mu_{kml} + \ln p(C_k)$. If we take $ln\mu_{kml} = A_{kml}$ (which are essentially constants) and $lnp(C_k) = B_k$, we have $a_k = \sum_{l=1}^{L}\sum_{m=1}^{M} A_{kml}x_{ml} + B_k$. As this represents a linear equation in $x_m l$, a_k is linear in the components of $\mathbf{x}$.

Here μ_{kml}, is the probability of seeing a 1 in the l^{th} position of the m^{th} feature while dealing with the k^{th} class. $\therefore \sum_{k=1}^{K}\sum_{m=0}^{M}\sum_{l=0}^{L} \mu_{kml} = 1$.

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: If a data point $\mathbf{x}$ as been assigned to class j, then the value $\sum_k L_{kj} p(y=k|\mathbf{x})$ is minimum. $\therefore$ for two classes (0 and 1), if fraction $\frac{\sum_k L_{k0} p(y=0|\mathbf{x})}{\sum_k L_{k1} p(y=1|\mathbf{x})} > 1$, the denominator (loss because of data point being assigned class C_1) has a smaller value than that of the numerator the point is assigned to C_1, else C_0. When the fraction is 1, it represents equal loss for both classes and hence a decision boundary.

$\therefore$ we have $\frac{\sum_j L_{k0} p(y=0|\mathbf{x})}{\sum_j L_{k1} p(y=1)|\mathbf{x})} = \frac{L_{00}p(y=0)|\mathbf{x} + L_{01}p(y=0)|\mathbf{x})}{L_{10}p(y=1)|\mathbf{x} + L_{11}p(y=1|\mathbf{x})} = \frac{L_{00}P(y=0)P(\mathbf{x}|y=0) + L_{01}P(y=0)P(\mathbf{x}|y=0)}{L_{10}P(y=1)P(\mathbf{x}|y=1) + L_{11}P(y=1)P(\mathbf{x}|y=1)} = 1$ for decision boundary. As $L_{00} = L_{11} = 0$, the equation boils down to $\frac{L_{01}P(y=0)P(\mathbf{x}|y=0)}{L_{10}P(y=1)P(\mathbf{x}|y=1)} = 1$. Since, $P(x|y=0) \sim \frac{1}{0.5\sqrt{2\pi}} e^{-2x^2}$, $P(x|y=1) \sim \frac{1}{\sqrt{0.5}\sqrt{2\pi}} e^{-x^2}$ and $P(y=1) = P(y=0) = 0.5$, we have $\frac{\sqrt{2} \cdot 2(0.5) e^{-2x^2}}{1 \cdot \sqrt{2}(0.5) e^{-x^2}} = 2e^{-x^2} = 1$. This implies $-x^2 = \ln(0.5)$ or $x = \pm\sqrt{\ln(2)}$.

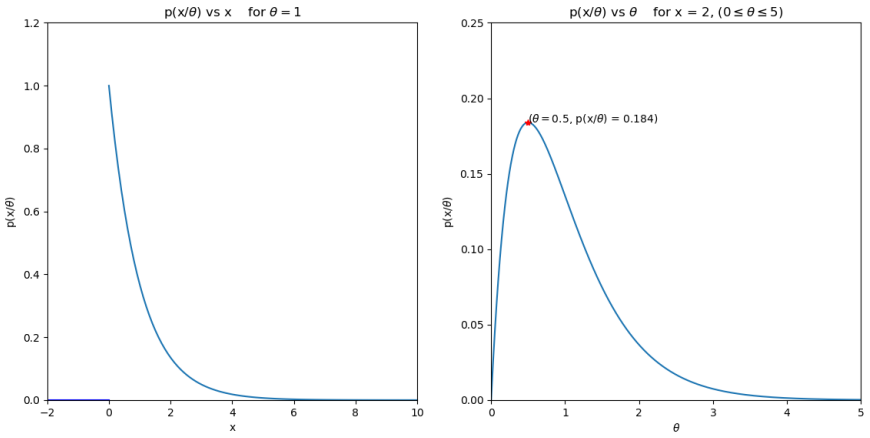
Hence we have two separating boundaries, i.e. $x = -\sqrt{\ln(2)}$ and $x = \sqrt{\ln(2)}$.

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:



(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: likelihood of θ given x is $\mathcal{L}(\theta|x) = p(x|\theta)$
Given that n samples are independent

$$\begin{aligned}\mathcal{L}(\theta|x_1 \cdots x_n) &= \mathcal{L}(\theta|x_1)\mathcal{L}(\theta|x_2)\mathcal{L}(\theta|x_3)\cdots\mathcal{L}(\theta|x_n) = \prod_{i=1}^{n}\mathcal{L}(\theta|x_i) \\ &= p(x_1|\theta)p(x_2|\theta)p(x_3|\theta)\cdots p(x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta) \\ &= (\theta e^{-\theta x_1})(\theta e^{-\theta x_2})(\theta e^{-\theta x_3})\cdots(\theta e^{-\theta x_n}) \\ &= \theta^n e^{-\theta(x_1+x_2+x_3\cdots x_n)} = \theta^n e^{-\theta(\sum_{i=1}^{n} x_i)}\end{aligned}$$

To find the "maximum likelihood" we have to take derivative of this and equate it to 0 to solve θ.

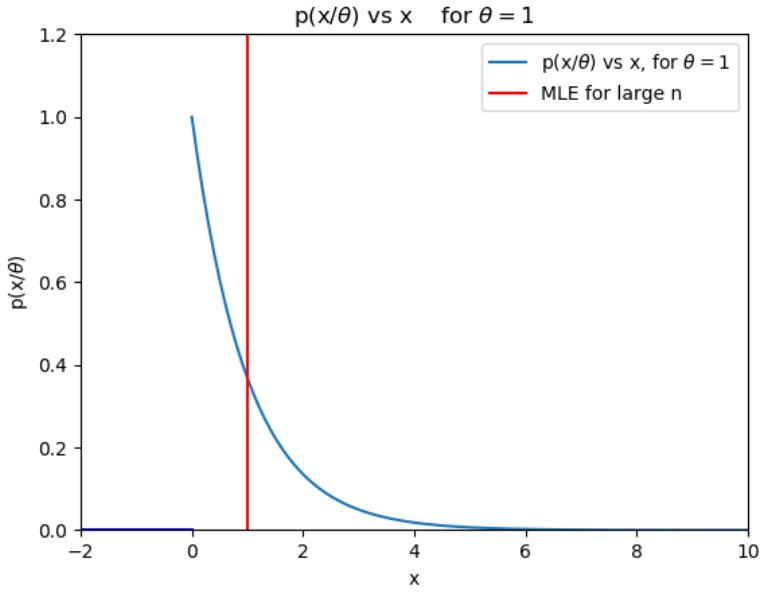
$$\frac{d}{d\theta}\left(\theta^n e^{-\theta(x_1+x_2+x_3\cdots x_n)}\right) = 0$$

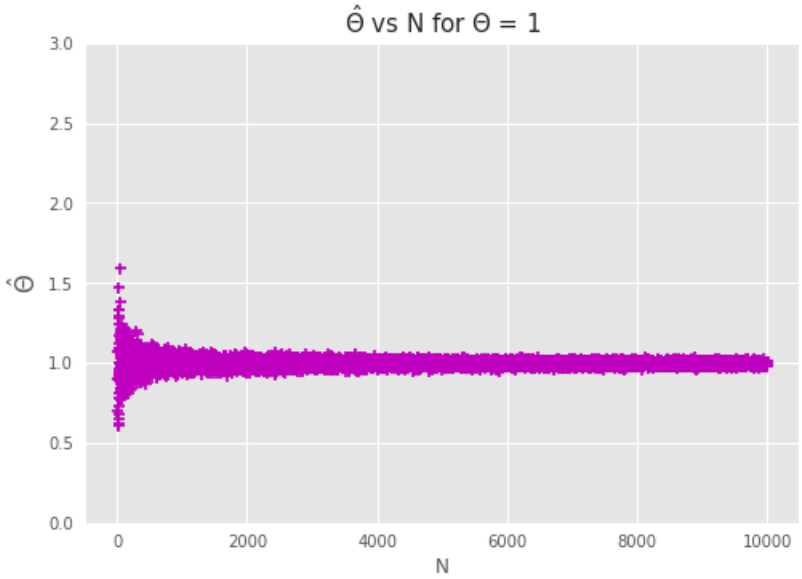
Taking log-likelihood derivative of this function, we get

$$
\begin{aligned}
&\implies \frac{d}{d\theta} log_e \left(\theta^n e^{-\theta(x_1+x_2+x_3\cdots x_n)}\right) = 0 \\
&\implies \frac{d}{d\theta} \left(log_e(\theta^n) + log_e(e^{-\theta(x_1+x_2+x_3\cdots x_n)})\right) = 0 \\
&\implies \frac{d}{d\theta} \left(n \cdot log_e(\theta) - \theta(x_1 + x_2 + x_3 \cdots x_n))\right) = 0 \\
&\implies \frac{n}{\theta} - (x_1 + x_2 + x_3 + \cdots x_n) = 0 \\
&\implies \frac{n}{\theta} = (x_1 + x_2 + x_3 + \cdots x_n) \\
&\implies \hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i}
\end{aligned}
$$

(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: To solve this, we use the Law of Large Numbers which states that for a very large sample ($n \to \infty$) of i.i.d variables, their expectation (or MLE) converges to the mean of the distribution. $\therefore \lim_{n\to\infty} \theta \to \mu$. $\mu = \int_0^\infty xp(x) = \int_0^\infty x\theta e^{-\theta x} = \theta(\frac{xe^{-\theta x}}{-\theta}|_0^\infty - \int_0^\infty \frac{e^{-\theta x}}{-\theta}) = \theta \cdot \frac{1}{\theta^2} e^{-\theta x}|_0^\infty = \frac{1}{\theta}$. $\therefore$ the MLE $\hat{\theta}$ approaches $\frac{1}{\theta} = 1$ for large n.





As seen in the plot above, the MLE (for $\theta = 1$) converges to 1 for very large n.

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:

$$\mathcal{L}(\alpha,\lambda|X) = \frac{\lambda^{\alpha}X^{\alpha-1}e^{-\lambda X}}{\Gamma(\alpha)}$$

Given that n samples are independent and identically distributed

$$\begin{aligned}
\mathcal{L}(\alpha,\lambda|X_1 \cdots X_n) &= \mathcal{L}(\alpha,\lambda|X_1)\mathcal{L}(\alpha,\lambda|X_2)\cdots\mathcal{L}(\alpha,\lambda|X_n) = \prod_{i=1}^{n}(\mathcal{L}(\alpha,\lambda|X_i)) \\
&= \prod_{i=1}^{n}\left(\frac{\lambda^{\alpha}{X_i}^{\alpha-1}e^{-\lambda X_i}}{\Gamma(\alpha)}\right) \\
&= \left(\frac{\lambda^{\alpha}{X_1}^{\alpha-1}e^{-\lambda X_1}}{\Gamma(\alpha)}\right)\left(\frac{\lambda^{\alpha}{X_2}^{\alpha-1}e^{-\lambda X_2}}{\Gamma(\alpha)}\right)\cdots\left(\frac{\lambda^{\alpha}{X_n}^{\alpha-1}e^{-\lambda X_n}}{\Gamma(\alpha)}\right) \\
&= \frac{\lambda^{n\alpha}e^{-\lambda\sum_{i=1}^{n}X_i}\prod_{i=1}^{n}X_i^{\alpha-1}}{(\Gamma(\alpha))^n}
\end{aligned}$$

Given that α is known, To find the MLE of λ we have to take derivative with respect to λ and equate it to zero.

$$\implies \frac{d}{d\lambda}\left(\frac{\lambda^{n\alpha}e^{-\lambda\sum_{i=1}^{n}X_i}\prod_{i=1}^{n}X_i^{\alpha-1}}{(\Gamma(\alpha))^n}\right)=0$$

Taking log-likelihood derivative of this function, we get

$$\implies \frac{d}{d\lambda}log_e\left(\frac{\lambda^{n\alpha}e^{-\lambda\sum_{i=1}^{n}X_i}\prod_{i=1}^{n}X_i^{\alpha-1}}{(\Gamma(\alpha))^n}\right)=0$$

$$\implies \frac{d}{d\lambda}log_e\left(\lambda^{n\alpha}e^{-\lambda\sum_{i=1}^{n}X_i}\prod_{i=1}^{n}X_i^{\alpha-1}\right)-\frac{d}{d\lambda}log_e\left((\Gamma(\alpha))^n\right)=0$$

deriving with respect to λ makes α constant $\implies \frac{d}{d\lambda}log_e\left((\Gamma(\alpha))^n\right)$ is zero

$$\implies \frac{d}{d\lambda}log_e\left(\lambda^{n\alpha}e^{-\lambda\sum_{i=1}^{n}X_i}\prod_{i=1}^{n}X_i^{\alpha-1}\right)=0$$

$$\implies \frac{d}{d\lambda}log_e(\lambda^{n\alpha})+\frac{d}{d\lambda}log_e(e^{-\lambda\sum_{i=1}^{n}X_i})+\frac{d}{d\lambda}log_e(\prod_{i=1}^{n}X_i^{\alpha-1})=0$$

deriving with respect to λ makes α constant $\implies \frac{d}{d\lambda}log_e(\prod_{i=1}^{n}X_i^{\alpha-1})$ is zero

$$\implies \frac{d}{d\lambda}log_e(\lambda^{n\alpha})+\frac{d}{d\lambda}log_e(e^{-\lambda\sum_{i=1}^{n}X_i})=0$$

$$\implies \frac{d}{d\lambda}(n\alpha)log_e(\lambda)+\frac{d}{d\lambda}\left((-\lambda)\sum_{i=1}^{n}X_i\right)log_e(e)=0$$

$$\implies \frac{n\alpha}{\lambda}-\sum_{i=1}^{n}X_i=0$$

$$\implies \frac{n\alpha}{\lambda}=\sum_{i=1}^{n}X_i$$

$$\implies \lambda=\frac{n\alpha}{\sum_{i=1}^{n}X_i}$$