Problem 2 : Bayes Rule

1. What is total probability of phone being defective?

Answer Given, A produces → 20 %, B → 30 % and C → 50%.

Then, Probability of phone being produced by factory A is

$$P(A) = 20/100 = 0.2$$

Similarly, $P(B) = 30/100 = 0.3$

$P(C) = 50/100 = 0.5$

Also given, probability of defective phone from A is 2%.

$$\Rightarrow P(D/A) = 2/100 = 0.02$$

Similarly, $P(D/B) = 1/100 = 0.01$

" $P(D/C) = \frac{0.05}{100} = 0.0005$

Now, To calculate the probability of a phone being defective, we can use law of total probability & Bayes Rule i.e. $P(A) = \sum_i P(A/B_i)\, P(B_i)$

$$\Rightarrow P(D) = P(D/A)\, P(A) + P(D/B)\, P(B) + P(D/C)\, P(C)$$

$$= (0.02) * (0.2) + (0.01) * (0.3) + (0.0005) * 0.5$$

$$\boxed{P(D) = 0.00725}$$

2. Probability that this defective phone is from factory A?

Answer To calculate this, we can use Baye's Rule :-

$$P(A/B) = \frac{P(B/A)\, P(A)}{P(B)}$$

Applying this in this problem :-

$$P(A/D) = \frac{P(D/A)\, P(A)}{P(D)}$$

Page No.

$$= \frac{(0.02) * (0.2)}{0.00725} = \frac{0.004}{0.00725}$$

$$\boxed{P(A/D) = 0.5517}$$

3. Probability that this defective phone is from factory B?

Answer Similarly, Using Bayes Rule -

$$P(B/D) = \frac{P(D/B)\ P(B)}{P(D)} = \frac{(0.01) * (0.3)}{0.00725}$$

$$\Rightarrow \boxed{P(B/D) = 0.4138}$$

4. Probability that this defective phone is from factory C?

Answer Similarly, again using Bayes Rule -

$$P(C/D) = \frac{P(D/C)\ P(C)}{P(D)} = \frac{0.0005 * 0.5}{0.00725}$$

$$\boxed{\Rightarrow P(C/D) = 0.03448}$$

Problem - 1 : Independence and law of Total probability

1. Compute $P(X=1)$, Given $P(Y=1) = 0.9$, $P(Z=1) = 0.8$, $P(X=1 \mid Y=1, Z=1) = 0.6$, $P(X=1 \mid Y=1, Z=0) = 0.1$, and $P(X=1, Y=0) = 0.2$. Y and Z are independent events.

Ans - Using law of total probability - $P(A) = \Sigma_i\ P(A/B_i)\ P(B_i)$

We have 3 cases here when $X=1$ i.e.

$$P(X=1 \mid Y=1, Z=1) = 0.6$$
$$P(X=1 \mid Y=1, Z=0) = 0.1$$
$$P(X=1, Y=0) = 0.2$$

Hence, applying law of total probability and summing up these 3 cases, we get →

Date __/__/____

Saathi

$P(X=1) = P(X=1 \mid y=1, z=1) \cdot P(y=1, z=1) +$
$P(X=1 \mid y=1, z=0) \cdot P(y=1, z=0) +$
$P(X=1, y=0) \cdot P(y=0)$.

Further, expanding $P(y=1, z=1)$ & $P(y=1, z=0)$ using conditional probability, we get →

$P(X=1) = P(X=1 \mid y=1, z=1) \cdot P(y=1) \cdot P(z=1) +$
$P(X=1 \mid y=1, z=0) \cdot P(y=1) \cdot P(z=0) +$
$P(X=1 \mid y=0) \cdot P(y=0)$
(as y and z are independent events).

$\Rightarrow P(X=1) = 0.6 * 0.9 * 0.8 +$
$0.1 * 0.9 * (1-0.8) +$
$0.2 * (1-0.9)$

$\Rightarrow$ **$P(X=1) = 0.47$**

2. Compute Expected value $E[y]$

Answer $E[y] = \Sigma (y * P(y=y))$, where summation is for all values of y.
here, we are $y=0$ or $y=1$

$\Rightarrow E[y] = 0 * P(y=0) + 1 * P(y=1)$
$= 0 * 0.1 + 1 * 0.9$
$E[y] = 0.9$

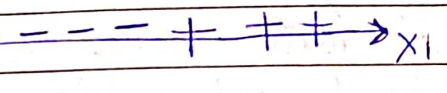
3. Given, instead of 0 and 1, y is taking 115 and 20 as values
$P(y=115) = 0.9$.

$\Rightarrow P(y=20) = 1-0.9 = 0.1$
$E[y] = 115 * P(y=115) + 20 * P(y=20)$
$E[y] = 105.5$

Date __/__/____

Saathi

Problem 3 : Feature Transformation & Kernels.

1. 

Consider 1-D dataset in Fig 1(a). Can you think of a 1-D transformation that will make the points linearly separable

Ans Yes, we can make the polynomial transformation of degree 2, which transform the points into a parabola. This following equation is for the same which will make these points linearly separable -

$$\phi(x) = x^2 \qquad \text{e.g. } \phi(x) = x^2 + c \ (c \text{ is constant})$$

2. Still consider the same above 1-D dataset (in Fig 1(a)). Can you think of a 2-D transformation that makes points linearly separable.

Ans Yes, we can have following 2-D transformation →

$$\phi(x, y) = [x, y^2 + c]$$

where x and y are variables & c is constant real value. The 1st dimension x is simply the original data points. The 2nd dimension y^2, is a quadratic function of the original data point. This means, the transformed data will be able to catch non-linear relationship between original data points.

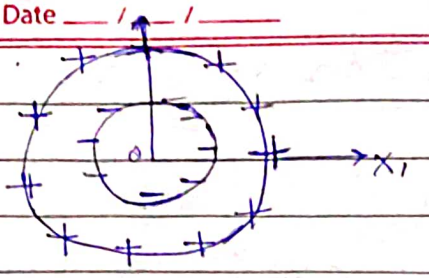
For e.g. → Lets say $X = [1, 2, 3]$

then, by applying $\phi(x, y) = [x, y^2 + 1]$, we get -

$x\_transformed = [[1, 2], [2, 5], [3, 10]]$

Date __/__/____

Saathi

3. 

Consider the 2-D dataset, can you suggest a 1-D transformation that will make the data linearly separable.

Ans Yes, we can use the following transformation technique that leverages the radial distance of each data point from the origin. Here are the steps-

a. Calculate the radial distance (r) of each data point from the origin $(0,0)$ in 2D space using the Euclidean distance formulae:

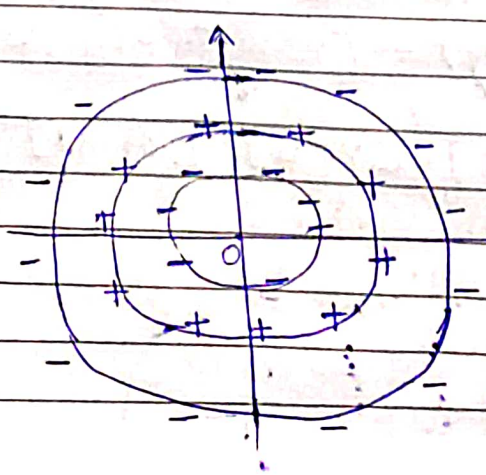
r = square root $(x^2 + y^2) = \sqrt{(x^2 + y^2)}$

b. Assign class label (1 and 0) to both classes respectively.

c. The transformed 1D dataset will consist of pairs of (r, label), where "r" = radial distance and 'label' is assigned class label (0 or 1)

d. we can use value r_threshold to classify datapoints as:- If $r <= r\_threshold$, then assign label 1.
& If $r > r\_threshold$, assign label 0.

By choosing appropriate 'r_threshold' and following these steps, we can achieve linear separation between two classes.

4. 

Can you suggest a 2D transformation of the dataset, that makes it linearly separable?

Answer

No, we cannot make datapoints belonging to different class linearly separable here in this case.

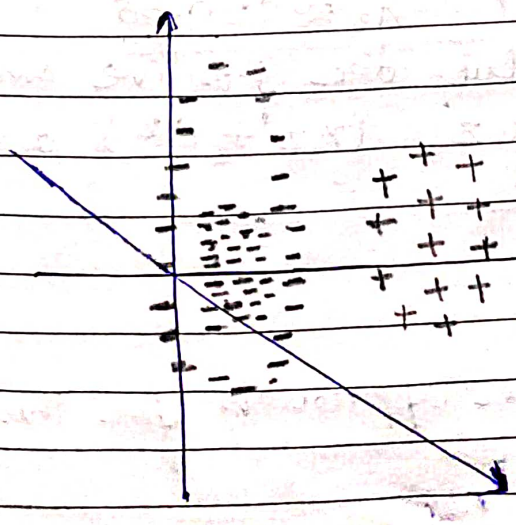
This figure has 3 concentric circles, the inner most and the outermost belongs to the same class, but the middle circle's datapoints belong to a different class.

The linear Kernel which is simple dot product of the input vectors essentially works in the original feature space and would not be suitable here where a transformation of dataset is required.

The RBF Kernel, which implicitly maps data into a higher-dimensional space, can potentially make the data linearly separable by transforming it into a high dimensional space where the classes become separable. This implies RBF will require to transform the dataset which is in 2-D in original space into a higher dimensional space.

Similarly, Polynomical Kernel is also incapable to make 2-D transformation to make classes linearly apart. Because, it also maps data into higher dimensional space using suitable polynomial functions.

For this dataset, a transformation from 2-D to 3-D is atleast required to make them linearly separable.



This figure depicts the 3-D transformed representation where negative and positive classes are linearly separable

$$\phi : x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in R^2 \rightarrow \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix} \in R^3$$

Date __/__/____

Kernel of Not : For the following functions, prove or disprove that it is a valid Kernel.

1. $K(x, z) = (xz + 1)^2$

Ans – To find if the function is a valid Kernel or not, we need to show Kernel matrix is symmetric and all of its eigen values are non-negative.

Symmetry :

$K(x, z) = (xz + 1)^2 = (zx + 1)^2 = K(z, x)$

$\therefore$, k is symmetric here.

For calculating eigen values, Lets construct a matrix of 3×3, $X = \{x_1, x_2, x_3\}$, $Z = \{z_1, z_2, z_3\}$

$$\Rightarrow K = \begin{vmatrix} (x_1 z_1 + 1)^2 & (x_1 z_2 + 1)^2 & (x_1 z_3 + 1)^2 \\ (x_2 z_1 + 1)^2 & (x_2 z_2 + 1)^2 & (x_2 z_3 + 1)^2 \\ (x_3 z_1 + 1)^2 & (x_3 z_2 + 1)^2 & (x_3 z_3 + 1)^2 \end{vmatrix}$$

$\Rightarrow$ Lets say $X = \{x_1 = 1, x_2 = 2, x_3 = 3\}$, $Z = \{z_1 = 1, z_2 = 2, z_3 = 3\}$

Substituting the values, we get →

$$\Rightarrow K = \begin{vmatrix} 4 & 9 & 16 \\ 9 & 25 & 49 \\ 16 & 49 & 100 \end{vmatrix}$$

$\Rightarrow$ eigen values $\lambda_1, \lambda_2, \lambda_3$ are :–

$\lambda \approx 0.15$, $\lambda_2 \approx 8.27$, $\lambda_3 \approx 120.58$

Since, all the eigen values are positive and K is symmetric, $K(x, z) = (xz + 1)^2$ is a valid Kernel

2. $K(x, z) = (xz - 1)^3$

Answer We will use the same approach for this 2nd part too.

Page No.

Symmetry:

$K(x,z) = (xz-1)^3 = (zx-1)^3 = K(z,x)$

therefore, K is symmetric

Now, calculating the kernel matrix, by substituting the same $x_1=1, x_2=2, x_3=3$ & $z_1=1, z_2=2, z_3=3$, we get -

$$K = \begin{vmatrix} (1.1-1)^3 & (1.2-1)^3 & (1.3-1)^3 \\ (2.1-1)^3 & (2.2-1)^3 & (2.3-1)^3 \\ (3.1-1)^3 & (3.2-1)^3 & (3.3-1)^3 \end{vmatrix}$$

$$\Rightarrow K = \begin{vmatrix} 0 & 1 & 8 \\ 1 & 7 & 26 \\ 8 & 26 & 64 \end{vmatrix}$$

Eigen values are $\lambda_1 \approx -3.56$, $\lambda_2 \approx 22.16$, $\lambda_3 \approx 45.40$

In this case, λ_1 is negative, means not all λs are non negative, which implies

$K(x,z) = (xz-1)^3$ is not a valid kernel for set of input points because they do not satisfy the Mercer's condition.

4. Problem 4 : Exponential Family & Geometric Distribution

1. Consider the geometric distribution parameterized by ϕ

$$p(y;\phi) = (1-\phi)^{y-1}\phi, \quad y = 1, 2, 3 \ldots$$

Show that geometric distribution is the exponential family, and give $b(y)$, η, $T(y)$ and $a(\eta)$.

Date __/__/____

Saathi

Proof:- To show that geometric distribution parameterized by ϕ belongs to the exponential family, we need to express its Probability mass Function (PMF) in the canonical form of exponential family. The PMF of geometric distribution is defined as:-

$$p(y;\phi) = (1-\phi)^{(y-1)}\phi$$

The canonical form of the distribution is given as-

$$p(y;\theta) = e^{(\eta(y) * T(y) - a(\eta(y)))}$$

where e = exponent,

$\eta(y)$ = the natural parameter

$T(y)$ = the sufficient statistic

$a(\eta)$ = the log partition function.

The value of $\eta(y)$, taking log of PMF, we get -

$$\eta(y) = \log(p(y;\phi)) = (y-1) * \log(1-\phi) + \log(\phi)$$

$$T(y) = y$$

Now, Lets put ϕ in terms of η

$$\phi = e^{(\eta(y) - (y-1) * \log(1-\phi))}$$

And $a(\eta)$ here will be log of sum of probabilities for all values of y and probability distributions sum to 1.

$$\Rightarrow a(\eta) = \log\left(\Sigma[y=1 \text{ to } \infty](1-\phi)^{(y-1)\phi}\right)$$

$$\Rightarrow a(\eta) = \log(\phi / (1-(1-\phi))) \quad \text{(This is geometric series with common ratio of } (1-\phi))$$

$$\Rightarrow a(\eta) = \log(\phi/\phi)$$

$$= \log 1$$

$$\Rightarrow \boxed{a(\eta) = 0}$$

or it can converges for infinite series:

$$a(\eta) = \log\left(\frac{1}{1-e^{\eta}}\right)$$

Page No.

Hence, putting values of $\eta(y)$, $T(y)$, $a(\eta)$, we can express the geometric distribution in the canonical form of the exponential family:-

$$P(y;\phi) = e^{(\eta(y) * T(y) - a(\eta))}$$

$$\boxed{P(y;\phi) = e^{((y-1) * \log(1-\phi) + \log(\phi) * y - 0)}}$$

2. Given a training set $\{(x_n, y_n)\}_{n=1}^{N}$ and let the log-likelihood of an example be $\log p(y_n | x_n; w)$. rule for learning using a GLM model with Geometric responses y.

Ans. To get the stochastic gradient ascent rule for learning GLM, we need following three assumptions —

1. Exponential Family Distribution – which we have got from 1st part of Problem 4. (Refer that).

~~$\eta = w^T x$~~

2. Link Function : linear predictor for the geometric function → $\boxed{g(\eta) = \log(\eta)}$

3. Linear predictor : $\eta = w^T x$

Log-Likelihood = Log [exp [$(y-1) \cdot \log(1-\phi) - \log(1 - \exp[\log(1-\phi)])$]]

$= (y-1) \cdot \text{Log}(1-\phi) - \log(1 - \exp[\text{Log}(1-\phi)])$

$= \eta^T (y-1) - \text{Log}(1 - \exp(\eta))$

$= (w^T x)^T (y-1) - \log(1 - \exp(\eta))$

Date ___/___/_____

Saathi

$$= (w^T x)^T (y-1) - \log(1 - \exp(w^T x))$$

$$= (x^T w)(y-1) - \log(1 - \exp(w^T x))$$

On computing gradient .w , we get -

$$\nabla w \log(y/x; w) = (y-1) * \nabla w \log(1-p) + \nabla w \log(p)$$

$$\Rightarrow \nabla w \log(1-p) = -\nabla w \eta$$

$$\nabla w \log(p) = \nabla w \eta$$

Substituting these values, we get -

$$\nabla w \log p(y/x; w) = (y-1)(-\nabla w \eta) + \nabla w \eta$$

$$\Rightarrow \quad = [(1-y) * \nabla w \eta] = (1-y) * x$$

applying Stochastic gradient ascent update rule →

$$w\_new = w\_old + \alpha * \nabla w \log p(y/x; w)$$

$$\boxed{w\_new = w\_old + \alpha(1-y)x \quad (\alpha = \text{learning rate})}$$

Moreover, we can also write

$$w + \alpha \left(\frac{\partial \log\text{-likelihood}}{\partial w} \right)$$

$$\boxed{= w + \alpha \left[x^T (y-1) - \frac{\exp(w^T x) \cdot x}{1 - \exp(w^T x)} \right]}$$