NPTEL Week-3 Live Session

on Machine Learning and Deep Learning - Fundamentals and Applications (noc24_ee146)

A course offered by: Prof. Manas Kamal Bhuyan, IIT Guwahati

NPTEL Quiz Solution: week-1, week-2

Powered by:

By

Arka Roy

NPTEL PMRF TA

Prime Minister's Research Fellow

Department of Electrical Engineering, IIT Patna

Web: https://sites.google.com/view/arka-roy/home

PMRF

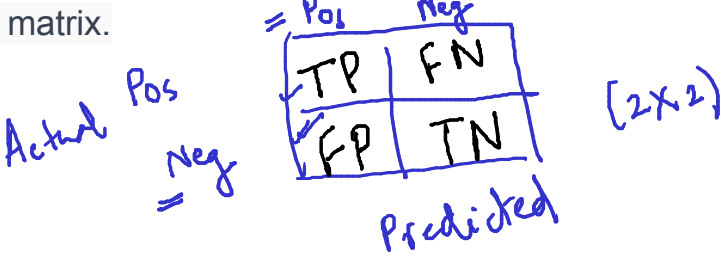
Prime Minister's Research Fellows

Ministry of Education

Government of India

1. In a binary classification problem, the confusion matrix is a _____ matrix.

- 1x1
- ~~2x2~~
- 3x3
- 1x2



2. Precision is defined as

- TP / (TP + TN)
- TP / (TP + FN)
- ~~TP / (TP + FP)~~
- TN / (TN + FP)

$$\text{Precision} = \frac{TP}{TP+FP}$$

3. In a binary classification problem, a classifier correctly predicts 90 instances as positive, incorrectly predicts 15 instances as positive when they are negative, correctly predicts 90 instances as negative, and incorrectly predicts 10 instances as negative when they are positive. What is the accuracy of the classifier?

FP = 15

FN

- 80
- 85
- ~~87.8~~
- 95

$\rightarrow TP = 90;\ FP = 15;\ TN = 90;\ FN = 10$

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} = \frac{90 + 90}{90 + 10 + 15 + 90} = 87.8\%$$

For the above question find the F1 score?

- 78.2%
- 85%
- ~~87.8%~~

$$F1\ score = \frac{2 \times Pre \times Rec}{Pre + Rec} = \left(\frac{2 \times 0.857 \times 0.9}{0.857 + 0.9}\right) = 0.878$$

Sensitivity

$$Rec = \frac{TP}{TP+FN} = \frac{90}{90+10} = 0.9$$

$$Prec = \frac{TP}{TP+FP} = \frac{90}{90+15} = \frac{90}{105} = 0.857$$

5) Consider a dataset with actual values (Y) and predicted values (Y_pred) given below:

Y = [5, 8, 12, 10, 15],

Y_pred = [4, 7, 10, 11, 13].

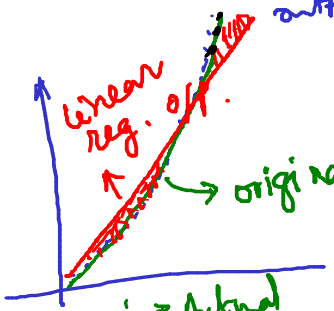
What is the bias of the model?

Linear regression :- weights

$\hat{y} = ax + b$ → bias

$\hat{y}$ → Prediction values.

$\{x_i, y_i\}$ — x_i: Input, y_i: expected Output / Actual values.



$Bias = E(y_{pred} - y_{true})$

$= Avg(y_{pred} - y_{true})$

$= Avg(-1, -1, -2, 1, -2) = (-1)$ Bias estimate

What is the variance of the model for the data given in the above question?

0

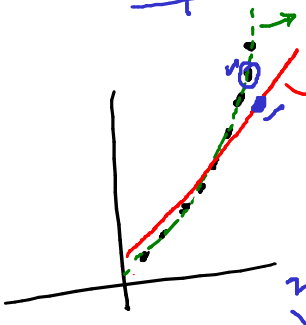
-1

2.2

None of the above

$E((x-\mu)^2)$

$E\begin{pmatrix}(x_1-\mu)^2\\(x_2-\mu)^2\\\vdots\end{pmatrix}$



$E\left((y_{pred} - y_{true})^2\right) = Avg(1, 1, 4, 1, 4)$

$= 2.2$

$= Avg\left((y_{pred} - y_{true})^2\right)$

Consider a dataset with actual values (Y) and predicted values (Y_pred) given below:

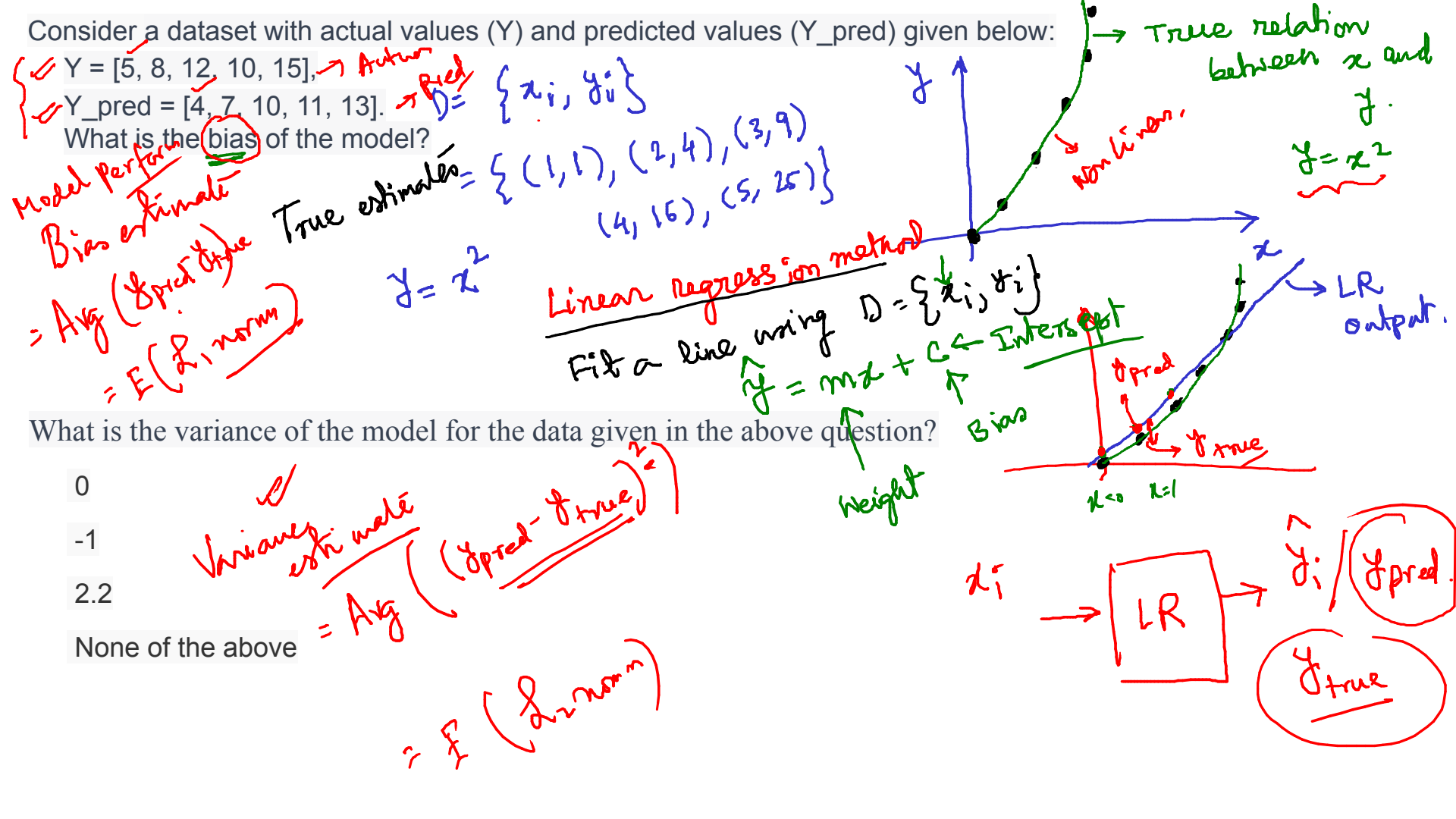
Y = [5, 8, 12, 10, 15],

Y_pred = [4, 7, 10, 11, 13].

What is the bias of the model?

What is the variance of the model for the data given in the above question?

- 0
- -1
- 2.2
- None of the above



Given X = {-2, -1, 0, 1, 2, 3, 4, 5, 6, 7} and the corresponding Y = {-0.5267, 1.3517, 3.8308, 5.5853, 7.5497, 9.9172, 11.2858, 13.7572, 15.7537, 17.3804}. Find the parameters of the linear regression model.

- 2.0065, 4.0312
- 2.0065, 3.5722
- 1.9214, 3.5722
- None of the above

$y + \epsilon$

$D = \{x_i, y_i\} \rightarrow$ LR $\rightarrow \hat{y} = wx + b = mx + c$

Learnable Parameters = $\{w, b\}$

$y = wx + b$ — (A)

$E(y) = E(wx + b)$

$E(y) = E(wx) + E(b)$

$\Rightarrow E(y) = wE(x) + b$

$\Rightarrow \left(\frac{1}{n}\sum y_i\right) = w\left(\frac{1}{n}\sum x_i\right) + b$

$8.58851 = 2.5w + b$ — (1)

Solve (1) & (2) :→ $w = 2.0065$, $b = 3.5722$

LR:-

$\hat{y} = 2.0065x + 3.5722$

$\hat{y}_1 (x=2) =$

$\hat{y}_2 (x=-1) =$

(A) × x on both side

$\Rightarrow xy = wx^2 + b$

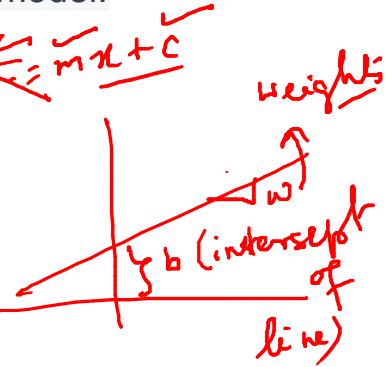
$E(xy) = wE(x^2) + E(b)$

$E(xy) = wE(x^2) + b$

$\Rightarrow \frac{1}{n}\sum x_i y_i = w\frac{1}{n^2}\sum x_i^2 + b$ — (2)

$\Rightarrow \sum x_i y_i = \frac{w}{n}\sum x_i^2 + nb$ — (2a)

$\Rightarrow 380.2522 = w \cdot 145 + 25b$ — (2c)



Find the MSE for the above question.

0.05783

0.04247

0.04876

None of the above

$$MSE = \frac{1}{n}\sum(y_{pred} - y_{true})^2$$

$$E\left(\left[y_{pred}\right] - \left[y_{true}\right]\right)^2 =$$

y_{pred}

$y = 2.0065x + 3.5722$ ..... }

$y_{pred} = \{-0.4408, \ldots\ldots$

$x = \{-2, -1, 0, 1, 2, 3, 4, 5, 6, 7\}$ }

$Y = \{-0.5267,$

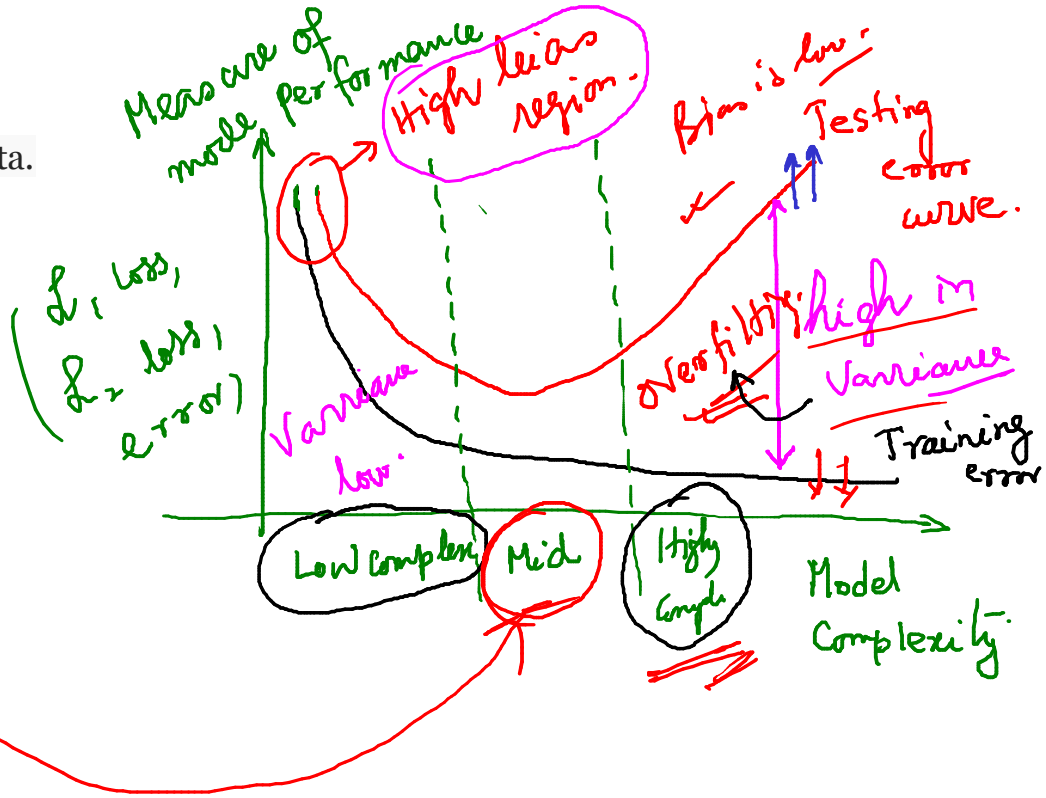
A model with high variance and low bias means

It can be too simple to understand the patterns of the data used in the training.

An excellent performance in the training data, but has a significant decrease in performance when evaluating the test data.

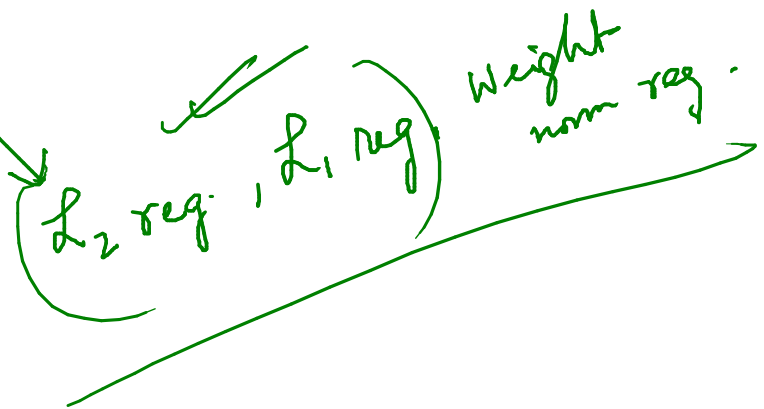
The model fits the test data better.

The model becomes less sensitive to the training data.



Which of the following techniques is used to prevent overfitting in machine learning?

- To create complex machine learning models.
- Train the model for more epochs. (you increase the training data).
- Using a regularization to the model.
- To increase the variance of the model.



Consider a binary classification problem with two classes, A and B with prior probability $P(A)=0.6$, $P(B)=0.4$. Let X be a single binary feature that can take values 0 or 1. Given: $P(X=1|A)=0.8$ and $P(X=0|B)=0.7$. Determine which class the classifier will classify when $X=1$

- Class A
- Class B
- Equiprobable for Class A and Class B
- Not enough information

Feature is distributed $P_{bernoulli}$

$P(A) = 0.6$

$P(B) = 0.4$

$X \in [0,1]$ ~ Bernoulli dist

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

class conditional

$P(X=1|A) = 0.8$

$P(X=0|A) = 0.2$

$P(X=0|B) = 0.7$

$P(X=1|B) = 0.3$

Posterior

$P(A|X) =$

$$P(A/X=1) = \frac{P(A).P(X=1/A)}{P(A).P(X=1/A) + P(B).P(X=1/B)} = \frac{(0.8 \times 0.6)}{(0.8 \times 0.6) + (0.4 \times 0.3)}$$

$$= \frac{0.48}{0.6} = 0.8.$$

$$P(B/X=1) = \frac{P(B).P(X=1/B)}{0.6} = \frac{0.4 \times 0.3}{0.6} = 0.2$$

$P(A/X=1) > P(B/X=1)$

$x \in A.$

Bayes' decision theory assumes that:

- The feature vectors are dependent on each other.
- The feature vectors are normally distributed.
- The feature vectors are identically distributed.
- The feature vectors are uniformly distributed.

Bayesian Theory

Discrimination function

$g(x)$

$g(x) = P(w/x)$

$x = \{x_1, x_2, x_3 \cdots x_d\}$

Multivariate

iid

$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$

Multivariate gaussian distribution

$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$ independent

$\Sigma = \begin{bmatrix} \cdots \\ \cdots \\ \cdots \end{bmatrix}$ Dependent

Assume that the word 'offer' occurs in 80% of the spam messages in my account. Also, let's assume 'offer' occurs in 10% of my desired e-mails. If 30% of the received e-mails are considered as a spam, and I will receive a new message which contains 'offer', what is the probability that it is spam?

- 0.778
- 0.774
- 0.668
- 0.664

$P(\text{offer}/\text{spam}) = 0.8$

$P(\text{offer}/\text{nonspam}) = 0.1$

$P(\text{spam}/\text{offer}) = ?$

$P(\text{spam}) = 0.3$

$P(\text{nonspam}) = 0.7$

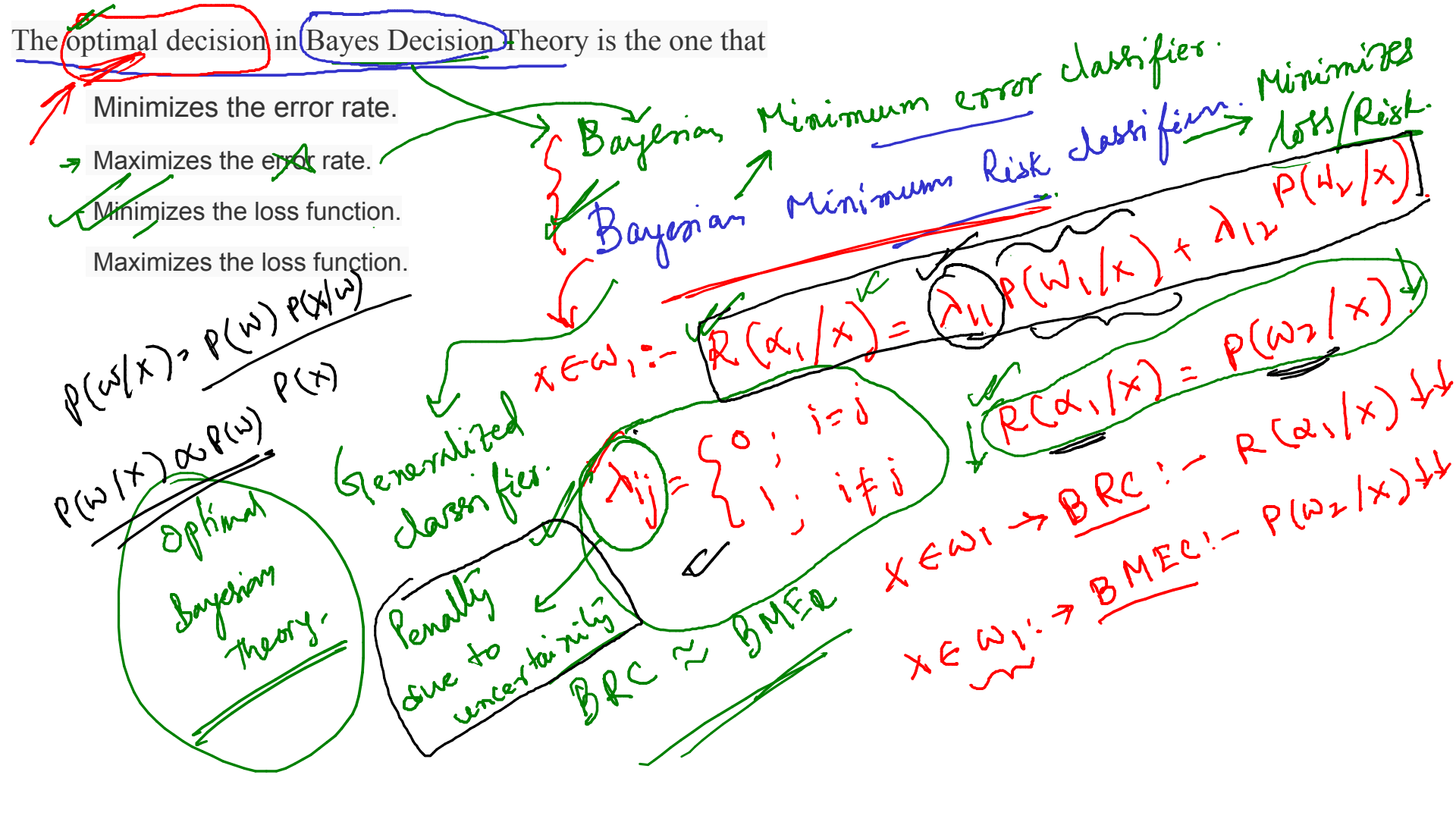
Posterior ← Prior, class cond.

$$P(\text{spam}/\text{offer}) = \frac{P(\text{spam}) \cdot P(\text{offer}/\text{spam})}{P(\text{spam}) \cdot P(\text{offer}/\text{spam}) + P(\text{nonspam})\, P(\text{offer}/\text{nonspam})}$$

$$= \frac{0.8 \times 0.3}{(0.8 \times 0.3) + (0.7 \times 0.1)} = 0.774$$

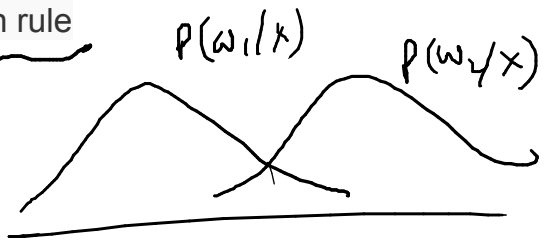
The optimal decision in Bayes Decision Theory is the one that

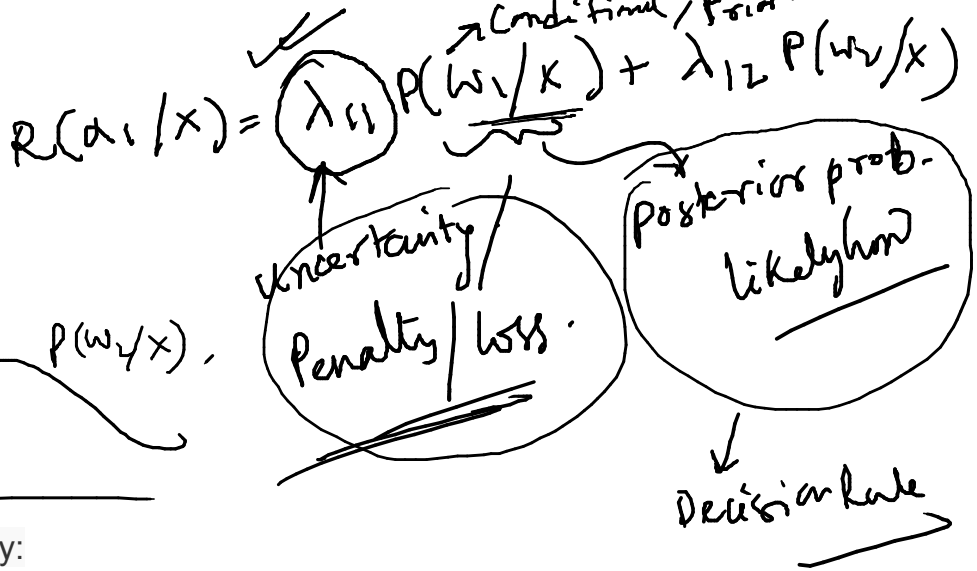
- Minimizes the error rate.
- Maximizes the error rate.
- Minimizes the loss function.
- Maximizes the loss function.



The risk function in Bayesian decision theory combines:

- The prior probabilities and the likelihood function.
- The decision boundaries and the feature vectors.
- The training set and the test set.
- The loss function and the decision rule





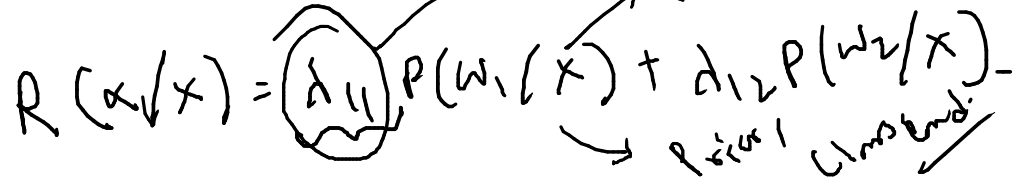
The loss function used in risk-based Bayesian decision theory:

- Quantifies the cost of different types of errors.
- Is equal to the likelihood function.
- Ignores the prior probabilities of the classes.
- Is not used in the decision-making process.

R ∝ λ.

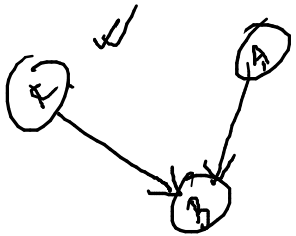
The risk-based Bayesian decision rule accounts for the consequences of different decisions by considering the:

- Number of features in the dataset
- The complexity of the classifier
- Uncertainty in the data and the associated losses
- Mean and standard deviation of the feature vectors



The generalized form of a Bayesian network that represents and solves decision problems under uncertain knowledge is known as an?

- Directed Acyclic Graph
- Table of conditional probabilities
- Influence diagram
- None of the above



Consider the following Bayesian network, where F = having the flu and C = coughing:

Determine the probability P(F|C) for the following Bayesian network so that it specifies the same joint probabilities as the given network.

0.23

0.03

0.35

None of the above.

$P(F) = 0.1$; $P(\bar{F}) = 1 - P(F) = 1 - 0.1 = 0.9$

$P(F/C) = ?$

$P(C/F) = 0.8$

$P(C/\bar{F}) = 0.3$

$$P(F/C) = \frac{P(F) \cdot P(C/F)}{P(F) \cdot P(C/F) + P(\bar{F}) \cdot P(C/\bar{F})} = \frac{0.1 \times 0.8}{(0.1 \times 0.8) + (0.9 \times 0.3)}$$

(Flu part)

$$= \frac{0.2286}{\approx 0.23}$$

For the above question, Are C and F independent in the given Bayesian network?

Yes.

No.

Can't say.

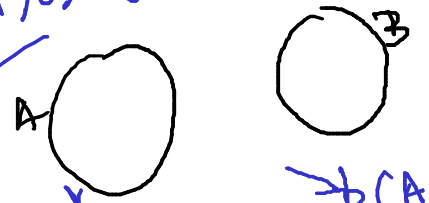
Insufficient information.

$P(F/C) = P(F)$ if C, F → independent

$P(F/C) \neq P(F)$

$P(F) = 0.1$

$P(F/C) = 0.23$

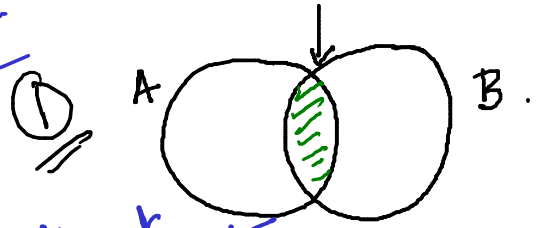


A & B are independent on each.

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

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

$P(A/B) = P(A)$



A & B are dependent events

$P(A/B) =$ Proportion of event A & B when they occur simultaneously / Probability of B.

Dependent

$P(A/B) = \frac{P(A \cap B)}{P(B)}$ → Bayes Rule