Lec: 07
02-07-24

Discriminant Analysis:

More than 2 dataset

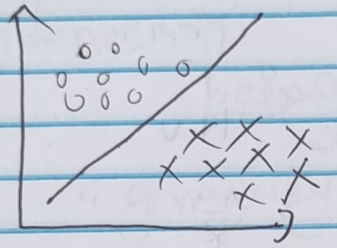
↓

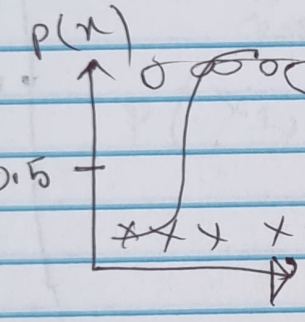
logistic regression is not good

2 classes → not ~~stable~~ → well separated data

↓

But not stable in estimating





> well separated classes → unstable

Generative Models:

→ Generate the function

↓

from the data.

→ Parametric model

↓

assumptn from data

↓

and then use Bayes theorem to flip probabilities

Bayes theorem → finding probability

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

ex:

Bayes theorem:

Email	X_1 length	X_2 contains "giveaway"	spam
Short	~~yes~~	Yes	Y
Long		NO	NO
Short		Y	Y
Long		N	N
Short		N	Y
Long		Y	N
Short		N	N
Long		Y	Y

$$P(\text{spam} \mid X_1, X_2) = \frac{P(X_1, X_2 \mid \text{spam}) \cdot P(\text{spam})}{P(X_1, X_2)}$$

i) calculate probability of $P(\text{spam}) = \frac{4}{8} = \frac{1}{2}$

(2) Likelihood of $p(x_1, x_2 \mid \text{spam}) = \frac{2}{4} = \frac{1}{2}$

x_1 = short
x_2 = Yes

~~$P(x_1, x_2 \mid \text{Not spam})$~~

(3) Evidence : The probability of email being short (OR) contain "giveaway" irrespective of being spam or not

$$P(x_1, x_2) = P(x_1, x_2 \mid \text{spam}) \times P(\text{spam}) + P(x_1, x_2 \mid \text{Not spam}) \cdot P(\text{Not spam})$$

$$= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2}$$

short yes yes

$$= \frac{1}{4} + \frac{1}{4} = 0.375$$

$$+ \frac{1}{4} = \frac{1}{2}$$

$$= \frac{1}{4} = 0.25$$

Chapter 4
* Bayes theorem

naive bayes

$$P(\text{spam} \mid X_1, X_2) = \frac{P(X_1 X_2 \mid \text{spam}) \times P(\text{spam})}{P(X_1, X_2)}$$

$$= \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{1}{4}} = 1$$

LDA : → (Parametric)

→ The data is of a uniform distribution (∩)

$$f(x) = \frac{1}{\sqrt{2\pi r}} e^{-1/2 \left(\frac{x-\mu}{\sigma}\right)^2}$$

» normal Distribution function

Assume fx → then put in naive bayes

but in LDA → putting this fx to the naive bayes

Bayes theorem → then putting normal distribut inside this → LDA (cutting)

Probability of $P_k(x) =$

$$\frac{\pi_k \times \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^{k} \pi_l \times \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(x/}}$$

QDA →

→ Parametric

→ Assumes the data has a quadratic form

→ Assumes all classes / Labels have their own covariance matrix.

→ why → not logistics? → naturally built for binary mod.

Generative model

↓

generate fn out of data

↓

data look like this → Predicting fn

↓

Then LDA → assumes ND-fn

↓

variant of LDA

↓

QDA (Assumes quadrat fn form

then flipping this into the bayes theorem.

Lec: 08
02-12-24

→ Logistic Regression
→ LDA
→ KNN
→ Naive Bayes classifier

Bayes Theorem:

$$P(Y|X) = \frac{P(X|Y) \times P(Y)}{P(X)}$$

Naive Bayes classifier:

i) All predictors are independent

conditional independence:

The effect of the value of a predictor (X) on a given class (Y) is independent of others predictors.

$P(Y)$ → prior probability
$P(X|C)$ → Likelihood
$P(X)$ → evidence.

$x_1, x_2 \dots x_p$

$$P(Y|X) = \frac{P(x_1|Y) \times P(x_2|Y) \times \dots \times P(x_p|Y)}{\text{likelihood}} \times \underbrace{P(Y)}_{\text{prior}}$$

#

free	money	class
1	0	Spam
1	1	Spam
0	1	Not spam
0	0	"
1	0	Spam
0	1	Not "
1	1	Spam
1	0	Spam

	Spam	Not spam
free	5	0
money	2	2
	7	2

word free appeared 5 times in spam

(1) calculate prior

$P(S) = \frac{5}{8}$ $P(NS) = \frac{3}{8}$

(2) calculate likelihood

$$p(\text{Free}|S) = \frac{\text{Number of free occurrences}}{\text{Total spam emails}}$$

$$= \frac{5}{5} = 1$$

$$p(\text{Money}|S) = \frac{\text{Number of money "}}{\text{Total spam emails}}$$

$$= \frac{2}{5}$$

$$p(\text{free}|NS) = \frac{\text{Number of free occ}}{\text{Total not spam emails}}$$

$$= \frac{0}{3} = 0$$

$$p(\text{money}|NS) = \frac{\text{Nur of money}}{\text{Total}}$$

$$= \frac{2}{3}$$

~~$p(Y|X)$~~

$\neq$ Test :

"free money click the link below"

$$P(S \mid \text{free}, \text{money}) = P(\text{free} \mid S) \times P(\text{money} \mid S) \times P(S)$$

$$= 1 \times \frac{3}{8} \times \frac{5}{8}$$

$$= \frac{3}{8} = \frac{1}{4} = 0.25$$ probability

$$P(NS \mid \text{free}, \text{money}) = P(\text{free} \mid NS) \times P(\text{money} \mid NS) \times P(NS)$$

$$= 0 \times \frac{2}{3} = 0$$

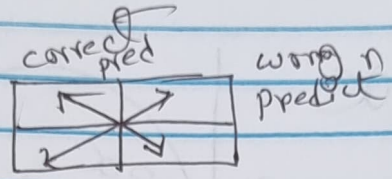
of free money as being the spam.

free credit card, free. . .

Confusion Matrix:

evaluat^n of classificat^n Model:

True positive → predictor + outcome =



How can we evaluate accuracy?

correct predict^n out of all dataset

$$\frac{\text{correct pred}}{\text{Total}}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision:

$$\frac{TP}{\text{Total positive}} = \frac{1}{2}\,\frac{TP}{TP+FP}$$

	Actual Positive	Actual Negative
Predicted Positive	2	3
Predicted Negative	4	2

Recall:

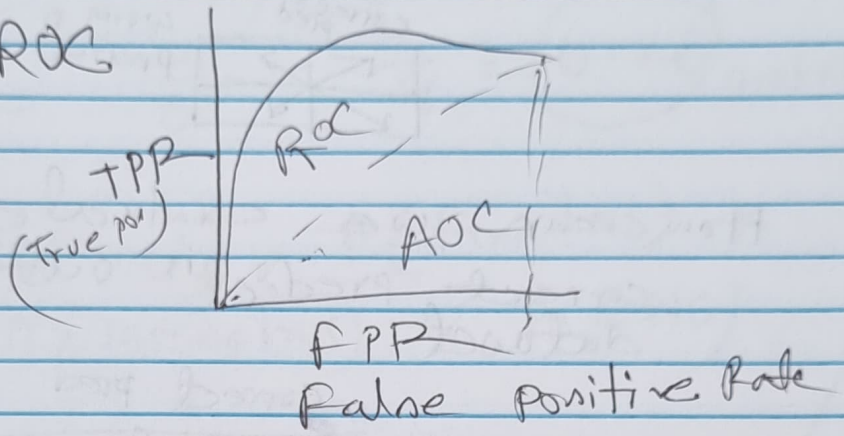
$$\frac{TP}{TP+FN}$$

F1-score → Mean / Avg for the precision and Recall

↓

closer to the accuracy

AUC-ROC



FPR → sould be 0

TPR → ↑↑

High TPR → than FPR doesn't matter

↓

Area should be bigger