2. Suppose we obtain a 'new' test $\langle \bar{x}_{n+1}, ? \rangle$

If $p(+|\bar{x}_{n+1}) > p(-|\bar{x}_{n+1})$ output '+'

otherwise output '−'

Qn:- what is $p(+|\bar{x}_{n+1}) \longrightarrow ?$

Bayes' Rule to the rescue:-

$$p(+|\bar{x}_{n+1}) = \frac{p(+, \bar{x}_{n+1})}{p(\bar{x}_{n+1})} = \frac{p(\bar{x}_{n+1}|+) \cdot p(+)}{p(\bar{x}_{n+1})}.$$

Similarly,

$$p(-|\bar{x}_{n+1}) = \frac{p(\bar{x}_{n+1}|-)\ p(+)}{p(\bar{x}_{n+1})}.$$

Comparison reduces to:- ~~$p(+|\bar{x}_{n+1}) \cdot p$~~ $p(x_{n+1}|+) \cdot p(+) \gtrless p(x_{n+1}|-) \cdot p(-)$?

In general, classes $C_1, C_2, \ldots, C_K$ (eg $K=10$ for digit recognition)
$\{+,-\}$, $\{0, \ldots, 9\}$.

Evaluate Find $\max_i \ p(C_i) \cdot p(\bar{x}_{n+1}|C_i) \longrightarrow$ output that.

Questions:- What is $p(C_i)$?

Prior belief about class C_i.

e.g., on 100 days, we played '30', maybe $p(\text{play}) \approx \frac{30}{100}$.

Say ~~30~~ empirically 30% of our emails are spam, then Prob(spam) = 0.3.

⊛ What are the features?

Who tells us to look @ weather, temp, ... before deciding to play, ... more philosophical.

Our features will be some high dimensional vector. we want nice representation.

" " Represent as a (high dimensional) vector"

Examples :-

① MNIST dataset 20×20, grayscale images.

each pixel $\in \{0, \ldots, 255\}$. / $\{0, 1\}$

Feature vector($\underline{x}$) = $(x_1, \ldots, x_{400})$, $x_i \in \{0, \ldots, 255\}$ / $\{0, 1\}$

is the ~~fea~~ value of pixel 'i'.

② News document / email. each word is a ~~value~~ feature

$\underline{x}$ - a length N document (has N words)
= $(x_1, \ldots, x_N)$, x_i - ith word.

Goal:- For each class estimate $P(\bar{x} \mid \text{class})$.

This is $P((x_1, \ldots, x_K) \mid \text{class})$.

- What are the challenges in estimating these probabilities (from samples).

Suppose feature i has $|F_i|$ F_i possible values.
e.g., MNIST, each pixel has 255/2 values.

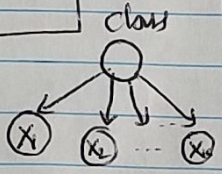
possible ~~images~~ feature vectors = $|F_1| \cdot |F_2| \cdots F_K$.

e.g., 255^{400} / 2^{400} ← very very large.

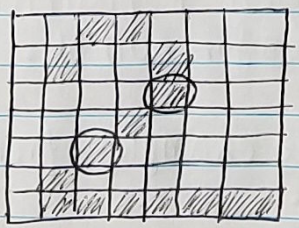
Same with text data.

Result from stat/prob:- To estimate M parameters needs 'M' samples. Impossible.

NAIVE BAYES assumption:- The features are "independent" given the class.



<very strong assumption>.



<next word is dependent on previous>.

$$P((x_1, \ldots, x_K) \mid \text{class} = c_i) = P(x_1 \mid \text{class} = c_i) \cdot P(x_2 \mid \text{class} = c_i) \cdots P(x_K \mid \text{class} = c_i).$$

how many parameters now?

$F_1 \cdots F_K \longrightarrow \boxed{(F_1 + F_2 + \cdots + F_K)}$.

$= K \cdot F$ if all F_i's are the same.

- 400 parameters for MNIST if on $\{0,1\}$.

documents/text :- $F \approx$ all words, $K \rightarrow$ 1000's.

Can make one more assumptions for text.
(BAG OF WORDS) model.
$P(x_1|\text{class}) = P(x_2|\text{class}) = \cdots = P(x_K|\text{class})$.

Assumptions :- All positions have same conditional distribution.
(BAD IDEA FOR IMAGES, WHY?)

- ~~Lee~~ Estimation of probabilities.

$$\arg\max_i \; P(C_i) \cdot P(\bar{x}_{n+1}|C_i)$$

$$= \max_i \; P(C_i) \cdot \prod_{j=1}^{K} P(x_{n+1,j}|C_i).$$

- ~~Binary~~ Take a feature 'i', with 'F' possible values.
eg, $F=2$, for binary, $F=3$ for outlook.

① $\longrightarrow \bar{x}(1), \ldots, \bar{x}(n) \longrightarrow$ all ~~data~~ examples from class C_i

② In these examples,
n_j = # examples with feature $i = j$.

n_0, n_1 = # examples with i^{th} feature $= \{0/1\}$

n_0, n_1 :- # examples with i^{th} feature $= \{0/1\}$.

Pick a class, c_i, and a feature, say the j^{th} feature x_j.

We want to estimate $P(\bar{x}_j = x \mid \text{Class is } c_i)$.

Ex:- MNIST, $c_i = 2$, $j = 192$, what is the distribution of the 192^{nd} pixel given the digit is 2.

Maximum Likelihood Estimates/ Empirical estimators:- (Bishop Chap 3) (Murphy Chap 3).

$$P(f(\bar{x}) = c_i) = \frac{\#\text{ examples with } f(\bar{x}) = c_i}{\#\text{ total examples}}.$$

$$P(\bar{x}_j = x \mid f(\bar{x}) = c_i) = \frac{\#\text{ examples with } \bar{x}_j = x \wedge f(\bar{x}) = c_i}{\#\text{ examples with } f(\bar{x}) = c_i}.$$

Why is it called MLE estimator?

- Insights into NB Classifiers :-

① NB assumption very strong, and does not hold in practice
- digit continuity.
still surprisingly good, why?

② Insufficient data / zero probability assignment.

- Suppose $x_j = x$, never happens in the training set.

X || eg, say digit = 2, pixel 192, never had '101' ($\in \{0, \ldots, 255\}$) but had...

Then, for the MLE estimator if the training example

has ~~er~~ j^{th} feature equal to 'x', $\underline{P(\bar{X}_{n+1} | C_j) = 0}$.

Straight away gone, no matter what else happens.

- One erroneous pixel / one wrong feature in test can make the probability zero.
  - Much more likely in words example.

~~Laplace~~ Laplace Smoothing. 'Smoothing of probabilities'.

(-) Suppose there are D choices/possibilities of a random variable, $D = 2$, coins, $D = 6$, die.

Add-β smoothing.

Suppose we have N outcomes, n_x = # x's in the outcomes

$$\text{MLE-Prob}(x) = \frac{n_x}{N}.$$

$$\text{Laplace-Prob}(x) = \frac{n_x + 1}{N + D}.$$ (Show this is indeed a probability).

In general, 'add-β' estimators:-

$$\text{Add-}\beta\ \text{Prob}(x) = \left(\frac{n_x + \beta}{N + D \cdot \beta}\right).$$

- Continuous random variables, Gaussian Naive Bayes.

Each continuous feature $\bar{X}_j$ is a Gaussian random variable, conditioned on the class.

eg, pixels, 

estimating distribution of a feature $\bar{X}_j$ under C_i,

(-) Take all the examples $\bar{X}_{(1)}, \ldots, \bar{X}_{(N_i)}$ that are in class C_i.

estimate :-

$$\text{Mean} = \frac{\bar{X}_j(1) + \ldots + \bar{X}_j(N_i)}{N_i} = \hat{\mu}_{j|i}$$

$$\text{Variance} = \frac{1}{N_i} \sum_{t=1}^{N_i} \left(\bar{X}_j(t) - \hat{\mu}_{j|C_i}\right)^2$$

(*) use logarithms for computation

Naive Bayes Classifiers

① Assumptions? / generative model

② Probability Estimation.

③ NB - assumption.

④

Tennis example :-

$P(\text{Class} = Y) = \frac{9}{14}$, $P(\text{No}) = \frac{5}{14}$.

Query $\boxed{\beta = \text{'1'}}$, $\boxed{F_{\text{Outlook}} = 3}$

$P(\text{Outlook} = S | \text{Class} = \text{No}) = \frac{3+1}{5+3} = \frac{1}{2} = \frac{4}{8}$ $\left(\frac{3}{5}\right)$

$P(\text{"} = R | \text{Class} = \text{No}) = \frac{2+1}{5+3} = \frac{3}{8}$ $\left(\frac{2}{5}\right)$

$P(\text{"} = O | \text{Class} = \text{No}) = \frac{0+1}{5+3} = \frac{1}{8}$ (0).

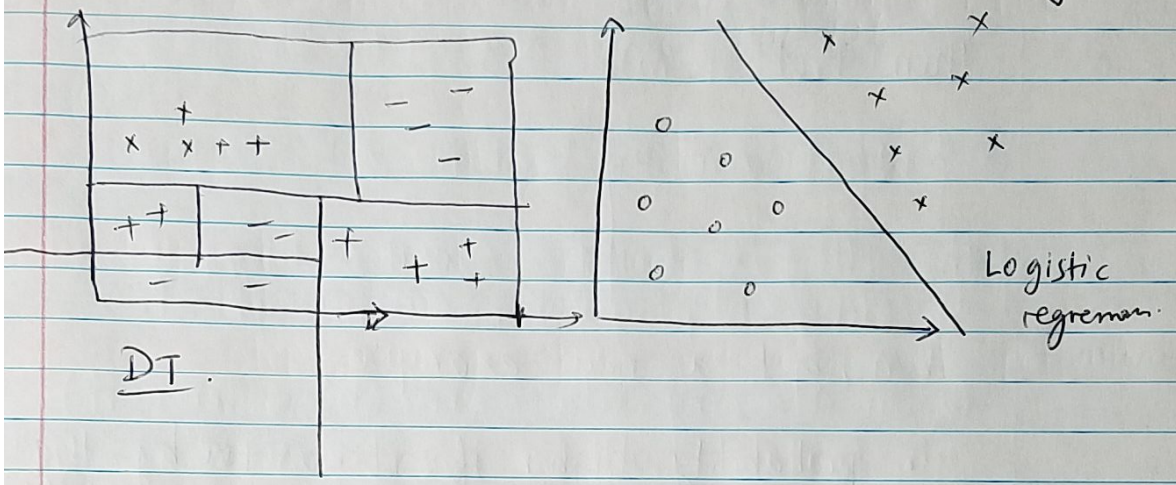
$P(\text{Humidity} = H | \text{Class} = \text{No}) = \frac{4+1}{5+2} = \frac{5}{7}$ $\left(\frac{4}{5}\right)$

$P(\text{Humidity} = N | \text{Class} = \text{No}) = \frac{1+1}{5+2} = \frac{2}{7}$ $\left(\frac{1}{5}\right)$

$P(\text{Sunny, High temp, Normal Humidity, ~~High~~ Strong Wind} | \text{Class} = \text{No})$

NB $= (\quad) = \left(\frac{4}{8} \cdot \frac{3}{8} \cdot \frac{2}{7} \cdot \frac{4}{7}\right)$.

Discriminative vs Generative Learning



Generative models (posi

breaking the "feature space" into regions.

(set of all possible features).

Ex:- 'N' → {Age} × {Boy?}
∈ $\mathbb{Z}_+$ {0, 1}.

Depending on where the data lies, we assign appropriate values.

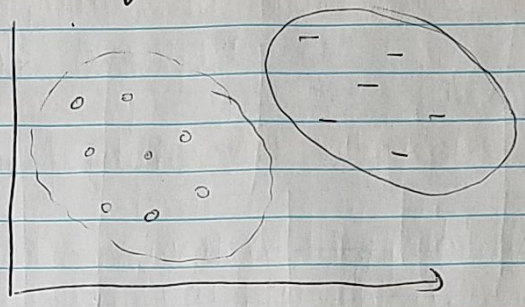
There is no 'model' for generating the examples.

Generative model - algorithm.

Look at one class at a time.
- say all +ve labels.
- Fit a probabilistic model to this part of the data.

⊛ detour a bit into estimating probabilites from data. (P → _).

Then do the same for -ve examples.



Then, get a new example,

- probability higher ~~if it is a~~ 'by the 'tve' model / '-ve' model?

Discriminative :- $\bar{x}$ (try to find) $f(\bar{x})$ (label).
Learn

Generative :- Learn $P(\bar{x}|f(\bar{x}))$ (generative model).
features labels.

Generative classification steps

⟨⊖ Find a generative model⟩

1. Using training examples "learn":-
   $P(\bar{x}|f(\bar{x}))$ and $P(f(\bar{x}))$
   ↓
   $P(\bar{x}|+)$ & $P(\bar{x}|-)$, $P(+)$ & $P(-)$.