Lecture 5 Naive Bayes (Mitchell 6.9, Murphy 3.5)

Recall:- If we know $p(\bar{x}, y)$, given $\bar{x}_{n+1}$

$$y^* = \arg\max_y P(y|\bar{x}) \qquad p(y) = \int_{\bar{x}} p(\bar{x}, y)\,d\bar{x}$$

$$= \arg\max_y p(y) \cdot \boxed{p(\bar{x}|y)}.$$

How to model $\hat{p}(y) \longrightarrow \dfrac{\#\text{examples w/ label } y}{\text{total \# examples}}$.

Naive Bayes:- $p(\bar{x}|y) = p((\bar{x}^1, \dots \bar{x}^d)|y) = p(\bar{x}^1|y) \dots p(\bar{x}^d|y)$

- Estimate each term individually
- avoids curse of dimensionality.

WHY is it an OK assumption?

Each attribute still gives info about labels.

EXAMPLE (WIKI) Label $\longrightarrow \{M, F\}$.

Features = (Height, Weight, foot size)

	Ht	Wt	Ft	Gender
1	6'	180	12	M
2	5'11"	190	11	M
3	5'7"	170	12	M
4	5'11"	165	10	M
5	5'	100	6	F
6	5'6"	150	8	F
7	5'5"	130	7	F
8	5'9"	150	9	F

NBA:- Height, Wt, Fot size are independent, given gender.

$\boxed{\text{Reasonable or not?}}$

$$Pr(H, W, Ft|M) = Pr(H|M) \cdot Pr(W|M) \cdot Pr(Ft|M)$$

$Pr(H,W,Ft|F) = Pr(H|F) \cdot Pr(W|F) \cdot Pr(Ft|F)$

Idea:- Each attribute, on its own gives some info abt labels and this is a simple way to aggregate them.

GOAL:- Estimate $\boxed{p(\bar{x}^i|y)}$ for all y, $\bar{x}^i$, $i=1,\dots,d$.

How to estimate one attribute's probabilities?

→ Probability Estimation

→ Inference algorithm.

MAP $y^* = \arg\max_y \underline{p(y|\bar{x})}$ ← Label most likely given data.

ML $y^* = \arg\max_y p(\bar{x}|y)$ ← label that explains data best.

ML:- find labels that maximize prob(observation).

Detour:- Probability Estimation

P• Given data/samples what is the underlying distribution?

Model:- A model class is a set of distributions that you pick from.

choose a model to explain the data.

Example:- 1. 6 coin tosses → $\boxed{HHTHTH}$

• same coin tossed independently '6' times.

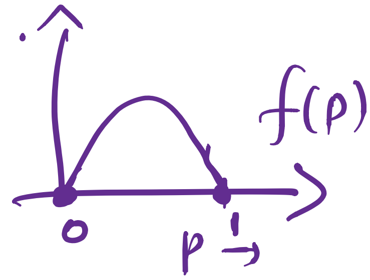
Model Class $\underline{\underline{\mathbb{P}}} = \{ \text{a coin with } p_H,\ \underline{\underline{p_H}} \in [0,1] \}$

$p_H \cdot p_H \cdot (1-p_H)\, p_H\, (1-p_H) \cdot p_H \cdot$

Principle of Maximum Likelihood (R. A. Fisher)

Pick the model that maximizes the probability of observed data.

'p' = prob(H), $\Pr(HHTHTH) = p^4(1-p)^2.$



$$\hat{p} = \arg\max_{p} \underbrace{p^4(1-p)^2} = f(p)$$

$$f'(p) = p^3(1-p)(4-6p) = 0 \qquad 4-6p=0 \Rightarrow \boxed{p = \frac{4}{6}}$$

'n' $\longrightarrow n_H,\ n_T$

then, $\hat{p}_{ML} = \dfrac{n_H}{n_H+n_T}$

$n_1 = \underline{5},\ n_2 = \underline{9},\ n_3 = \underline{12},\ n_4 = 4,\ n_5 = 3,\ n_6 = 3$

$$\max_{\substack{p(1)+\ldots\\+p(6)=1}} p(1)^5 \cdot p(2)^9 \ldots p(6)^3 \longrightarrow p(i) = \frac{n_i}{36} \quad \boxed{\hat{p}(1) = \frac{5}{36}}$$

$$\hat{p}(\bar{x}^i \mid y) = \frac{\#\text{examples where attribute is } \bar{x}^i,\ \&\ \text{label is } y.}{\#\text{examples w/ label } y}$$

Look at all examples with label 'y'.

Tennis.

The reminder of class was with annotated slides.

Key Point :-

By NB assumption, we need to model $p(\bar{x}^j|y)$,

i.e., For each label $y \in \mathcal{Y}$, model the distribution of the j^{th} attribute given label is 'y'.

How to do this?

1. Consider all examples with label 'y'.
2. consider the value of attribute 'j' for these examples.

	Ht	wt	Ft	Gender
1	6'	180	12	M
2	5'11"	190	11	M
3	5'7"	170	12	M
4	5'11"	165	10	M
5	5'	100	6	F
6	5'6"	150	8	F
7	5'5"	130	7	F
8	5'9"	150	9	F

In this example, suppose we want to model Heights given gender is female,

1. consider only examples 5, 6, 7, 8,
2. and the heights for these 4 examples.

Suppose we use 'GNB' (Gaussian Naive Bayes) to model heights, then the '4' green circled numbers are the numbers used to model $f(\underline{\text{Height}} \mid F)$.

EXERCISE :- convert heights to 'cm', then find the mean & std. deviation of resulting Gaussian we get.