Lecture 4.

* k- Nearest Neighbors (Recap Wrap-up).
* Naive Bayes (Probability ☺).

Goal :- figure map from features to labels

$$x \longrightarrow y$$

Discriminative model :- $x \rightarrow y$.

Generative model :- We try to model the entire process of generation of features & labels.

$$\boxed{p(\bar{x}, y)} \longrightarrow$$

$$\{(\bar{x}_1, y_1), \ldots, (\bar{x}_n, y_n)\} \longleftarrow \text{'}S\text{'}.$$

using 'S' $\longleftarrow$ $\boxed{p(\bar{x}, y)}$ $\longrightarrow$ distribution on (x, y).

Imagine, this is the "true process" by which each $(\bar{x}_i, y_i)$ is generated.

Given $\bar{X}_{n+1}$, what is the "best" y_{n+1} to predict?

$$y^* = \arg\max_y p(y|\bar{x}_{n+1})$$

OPTIMAL!! (MAP).

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

$$p(\bar{x}_{n+1}) = \sum_{y \in \mathcal{Y}} p(\bar{x}_{n+1}, y). \quad \text{(Marginalization)}$$

$$p(\bar{x}_{n+1}) \leftarrow$$

$$\max_y p(y \mid \bar{x}_{n+1}) = \max_y \frac{p(x_{n+1}, y)}{p(\bar{x}_{n+1})}$$

$$\arg\max_y p(y \mid \bar{x}_{n+1}) = \arg\max_y p(\bar{x}_{n+1}, y)$$

$$= \arg\max_y \boxed{p(y) \cdot p(\bar{x}_{n+1} \mid y)}$$

How do we learn $p(\bar{x}, y)$ from data(s)?

$$p(\bar{x}, y) = \underbrace{p(y)} \cdot \underbrace{p(\bar{x} \mid y)}$$

Learn these individually.

- Tennis :- $\mathcal{Y} = \{Y, N\}$, 9Y, 5N

$$\hat{p}(N) = 5/14$$

$$\hat{p}(Y) = 9/14$$

$\boxed{\text{WHY}}$ ...

$$\rightarrow \hat{p}(y) = \frac{\#\text{ example w/ label } y}{\#\text{ total examples}}.$$

- How to estimate $p(\bar{x}|y)$.
  - ROADBLOCK :- $\mathcal{X}$ could be large.

MNIST :- $|\mathcal{X}| = 256^{784} \ggg 60K$

TENNIS :- $|\mathcal{X}| = 3 \times 3 \times 2 \times 2 = \boxed{36}$

- In an ideal world :-

$$\hat{p}(\bar{x}|y) = \boxed{\frac{\#\text{examples } (\bar{x}, y)}{\#\text{examples w/ label } y}} = \frac{\hat{p}(\bar{x}, y)}{\hat{p}(y)}$$

MNIST :- $256^{784} \longrightarrow 256^{784}$

$\boxed{784}$ $(256) \longrightarrow \boxed{784 \times 256}$

If I give 256 images

- Each $\bar{x}_i$ — 'd' -dimensional feature vector.

"Perhaps", each dimension can be estimated separately!!

$$\left.\begin{array}{l} p(\text{wind} = w \mid y) \\ p(\text{wind} = s \mid y) \end{array}\right\} \longrightarrow \frac{\#\text{ex w/ weak wind \& Yes label}}{\#\text{ex w/ Yes label.}}$$

NAIVE BAYES ASSUMPTION :-

$\bar{x} = (\bar{x}^1, \bar{x}^2, \ldots, \bar{x}^d)$

The attributes are independent of each other, given the label.

$$p(\bar{x} \mid y) = p((\bar{x}^1, \ldots, \bar{x}^d) \mid y) \quad y = \text{'3'}$$

$$\overset{\text{NBA}}{=} p(\bar{x}^1 \mid y) \cdot p(\bar{x}^2 \mid y) \ldots p(\bar{x}^d \mid y)$$

Estimate each term separately!!!

Q :- How good is this assumption?