Lecture 12

เลื่อนสอบ ๓ กลุ่ม ป่วย
กรอกเอาไว้ ๕ นาที

Generative Learning

$$\underset{\sim}{D} \sim P(\underset{\sim}{X}, \underset{\sim}{Y})$$

r.v.s

generate

$$P(X, Y ; \underset{\sim}{\theta})$$

MLE (Frequentist)

$$\theta = \underset{\theta}{\operatorname{argmax}}\ P(D \mid \underset{\sim}{\theta})$$

parameter

Which θ maximizes the probability of seeing the dataset D

MAP (Bayesian)

$$\theta = \underset{\theta}{\operatorname{arg\,max}}\ P(\underset{\sim}{\theta} \mid D)$$

r.v.

Which θ is the most likely, given that we have observed the data set D

Summary of MAP

- If $n \rightarrow +\infty$, $\theta_{MAP} \longrightarrow \theta_{MLE}$.
- If n is small, the estimate will depends on the prior belief.
- MAP is a great estimator if an accurate prior belief is available.

True Bayesian Approach

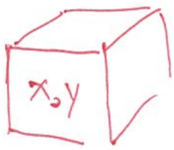
$$P(Y=y \mid X=x, D) = \underbrace{\int_{\theta} P(Y=y \mid \theta)\, P(\theta \mid D)\, d\theta}_{\text{average out all possible value of } \theta}$$

Bayes Classifier: return $\underset{\forall y}{\arg\max}\ P(Y=y \mid X=x;\theta)$

Estimating the conditional probability $P(Y=y \mid X=x)$

- Estimating $P(X=x, Y=y)$ is fine. But why don't we estimate $P(Y=y \mid X=x)$ directly.

Note: If we have enough data, we could estimate $P(X,Y)$ where we imagine a gigantic die that has one side for each possible value of (x,y).

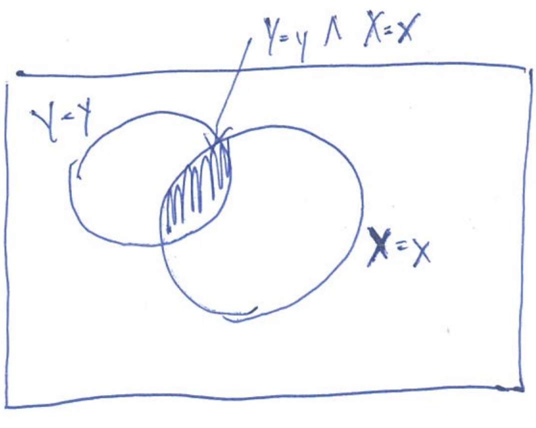


$P(X=x \wedge Y=y)$ is the probability that one specific side coming up.

Ex: By assuming X,Y together forms a R.V. that follows the binomial distribution, then we have the following by MLE

$$P(X=x \wedge Y=y) = \frac{\sum_{i=1}^{n} I(x_i=x \wedge y_i=y)}{n}$$

$$\underbrace{I}_{\text{indicator r.v.}}(x_i=x \wedge y_i=y) = \begin{cases} 1 & \text{if } x_i=x \text{ and } y_i=y \\ 0, & \text{otw.} \end{cases}$$



$$P(Y=y \mid X=x) = \frac{P(Y=y \wedge X=x)}{P(X=x)}$$

$$\boxed{\Rightarrow P(Y=y \mid X=x) = \frac{\dfrac{\sum_{i=1}^{n} I(x_i=x \wedge y_i=y)}{\cancel{n}}}{\dfrac{\sum_{i=1}^{n} I(x_i=x)}{\cancel{n}}}}$$

What's the problem?

- In d-dimensional space,

$$P(Y=y \mid \mathbf{X}=\underset{\sim}{x}) = P(Y=y \mid [\mathbf{X}]_1 = x_1, [X_2] = x_2, \dots, [X_d] = x_d)$$

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad \text{// feature vector}$$

$X = \{[X_1], [X_2], \dots, [X_d]\}$, Each X_i is r.v. associated w/ each coordinate

Practical Problem: Estimation by MLE is only good if we have many training vectors w/ same identical features as x. As $d \to +\infty$, this will never happens (e.g. image data)

$\Rightarrow$ If $d \to +\infty$, $P(Y=y, X=x) \to 0$ and $P(X=x) \to 0$

Naive Bayes Classifier:

- By Bayes rule, we have that $P(Y=y \mid X=x) = \dfrac{P(X=x \mid Y=y) \cdot P(Y=y)}{P(X=x)}$
- Again, we are in the world of generative learning
  - $\Rightarrow$ We have already know how to estimate $P(X=x)$ and $P(Y=y)$
  - $\Rightarrow$ How about estimating $P(X=x \mid Y=y)$??

Naive Bayes Assumption: All feature values are independent, given the label

$$P(X=x \mid Y=y) = \prod_{j=1}^{d} P([X_j] = x_j \mid Y=y)$$

- With the assumption, we can derive

$$h(x) = \arg\max_{\forall y} P(Y=y \mid X=x)$$

$$= \arg\max_{\forall y} \frac{P(X=x \mid Y=y)\, P(Y=y)}{P(X=x)}$$

$\leftarrow$ constant (unchange)

$$= \arg\max_{\forall y} \left(\prod_{j=1}^{d} P([X_j]=x_j \mid Y=y)\right) \cdot P(Y=y)$$

$$= \arg\max_{\forall y} \left[\sum_{j=1}^{d} \log\big(\underbrace{P([X_j]=x_j \mid Y=y)}_{\text{easy as one dimension}}\big) + \log(P(y))\right]$$

Ex: Spam filter by Naive Bayes / text classification

- Each vocabulary is one feature dimension
- We encode each email as a feature vector $x \in \{0,1\}^{|V|}$
- Each $x_j = 1$ iff the vocabulary x_j appears in the email)

$$\boxed{\text{email}} \longrightarrow x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \begin{array}{c} a \\ \vdots \\ z \end{array}$$

- $Y \in \{\text{SPAM} / \text{NOT-SPAM}\}$

- For a test email x_t, we would like to determine

$$P(Y=\text{SPAM} \mid X=x_t) \quad \text{and} \quad P(Y=\text{NOT-SPAM} \mid X=x_t)$$

$$- P(Y=\text{SPAM} \mid X=x_t) = P(Y=\text{SPAM} \mid [X_1]=x_1, [X_2]=x_2, \dots, [X_d]=x_d)$$

$$= \frac{\left(\prod_{j=1}^{d} P([X_j]=x_j \mid Y=\text{SPAM})\right)\left(P(Y=y)\right)}{P(X=x)}$$

In next lecture, we explore how to estimate this term

Visualization: