5 Naive Bayes

Last Time:

Data is drawn iid from an unknown distribution $D \sim P(X,y)$

We don't know what the unknown distr is

We estimate $P(X,y)$ by some distr dependent on parameter θ. $P_\theta(X,y)$

How to estimate θ?

MLE $\hat{\theta} = \arg\max_\theta \mathbb{P}(D)$ Frequentist

MAP $\hat{\theta} = \arg\max_\theta \mathbb{P}(\theta|D)$ Bayesian

How to make prediction once given $\hat{\theta}$?

Calculate $\mathbb{P}\{y|\theta\}$

More generally...

$$\mathbb{P}\{y|X=x\} = \int_0^\theta \underbrace{\mathbb{P}\{y|\theta\}}_{\text{Given a model what is prob it is } y} \underbrace{\mathbb{P}\{\theta|D\}}_{\text{prob of choosing a specific model}} d\theta$$

Naive Bayes

Given $P(X)$. $D \sim P(Y,X)$

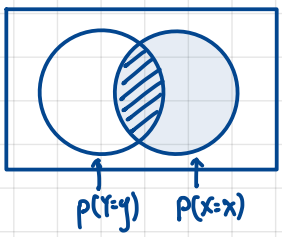
What is $P(Y=y|X=x)$

$$P(Y=y|X=x) = \frac{P(Y=y, X=x)}{P(X=x)}$$

In discrete cases...

$$= \frac{\sum_{i=1}^n I(X_i = x, Y_i = y)}{\sum_{i=1}^n I(X_i = x)}$$

(In sample estimate)



Basically, you discard all data where $X_i \neq x$

Hard to achieve if you have a large # of features

i.e. image. Requiring $X_i = x$ means all pixels must have the same value

Naive Bayes to rescue

$$\mathbb{P}(Y=y|X=x) = \frac{\mathbb{P}(X=x|Y=y)\,\mathbb{P}(Y=y)}{\mathbb{P}(X=x)}$$

$\mathbb{P}(Y=y)$. Not a problem. Few labels. Easy to est.

$\mathbb{P}(X=x)$. Normalizing

$\mathbb{P}(X=x|Y=y)$. Hard to estimate

Naïve Bayes makes crucial assumption

Given a label, the features are independent from each other

$$\mathbb{P}(X = \vec{x} \mid Y = y) = \prod_{i=1}^{d} \mathbb{P}(X_i = x_i \mid Y = y)$$

In the example of spam filtering

Two typewriters: One correspond to spam, other non-spam

At first, you decide whether to write a spam or non-spam

Once the decision is made, randomly strike the keys to compose email

Recall Bayes's classifier

$$h(\vec{x}) = \arg\max_y P(y \mid \vec{x}) = \arg\max_y \frac{P(x \mid y) P(y)}{Z}$$

constant doesn't matter

try to maximize

$$= \arg\max_y P(y) \prod_i P(x_i \mid y)$$ take log

$$= \arg\max_y \log(P(y)) + \sum_i \log(P(x_i \mid y))$$