WEEK 2

NAIVE BAYES

BAYESIAN CLASSIFICATION

→ is part of a family of **probabilistic** classification

- First, compute the bayesian probability of each class

$$P(y|x) = \frac{P(x|y)\,P(y)}{\sum_{y'} P(x|y')\,P(y')} = \frac{P(x|y)\,P(y)}{P(x)}$$

$P(x|y)$: class model; $P(y)$: Prior; $P(x)$: Normalizer factor

- Then the most probable class given observation

$$\hat{y} = \arg\max_{y} P(y|x)$$

COMPONENTS

- i.e. y ... UK patient has Ebola
  x ... observed symptoms

$P(y)$: prior probability of each class

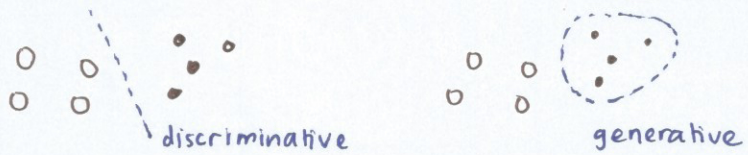
$P(x|y)$: class-conditional model
↳ describes how likely to see observation x for class y

$P(x)$: normalize probabilities across observations
↳ does not affect which class is most likely (arg max)
↳ makes probabilities more comparable to other data points / instances in your data

If we present a Bayesian classifier w/ an instance that is very unlike any instance it trained on, the value of the normalizer $\sum_y P(x|y)P(y)$ will be small

NAIVE BAYES: A GENERATIVE MODEL

→ Computes its predictions by modelling each class (not just by looking at boundary)



- $P(y|x) \propto P(x|y)\,P(y)$

→ All generative classifiers are probabilistic, BUT not all probabilistic classifiers are generative

- It is possible to ~~compute~~ estimate $P(y|x)$ directly
  ↳ i.e. logistic regression

→ For a probabilistic classifier, the decision boundary represents the points where classes are equally probable

INDEPENDENCE ASSUMPTION

→ We assume observations $x_1 \dots x_d$ are **conditionally independent** given y

$$P(x|y) = P(x_1, x_2, \dots x_d \mid y)$$
$$= [P(x_1 \mid x_2, \dots, x_d, y)\, P(x_2 \mid x_3, \dots, x_d, y) \dots P(x_d|y)]$$

CHAIN RULE

SIMPLIFIED TO

INDEPENDENCE

$$P(x|y) = P(x_1|y)\, P(x_2|y) \dots P(x_d|y)$$
$$= \prod_{i=1}^{d} P(x_i|y)$$

CONDITIONAL INDEPENDENCE

- 2 random events A & B are conditionally indep. given a third event C if the occurence of A and occurence of B are indep. events in their conditional probability distribution given C.

  i.e. Probabilities of going to the beach and getting a heat stroke are NOT independent: $P(B, HS) > P(B)\,P(HS)$

  ↳ May be independent if we know the weather is hot

  $$P(B, HS \mid hot) = P(B|hot)\, P(HS|hot)$$

  ↳ Hot weather 'explains' all the dependence between beach & heatstroke

- In classification, class value explains all the dependence between attributes

GAUSSIAN NAIVE BAYES CLASSIFIER [For real-valued data]

→ Suppose we're trying to distinguish children from adults based on size

- Classes: {a, c}
- Attributes: {height (cm), weight (kg)}
- Training examples: $\{h_i, w_i, y_i\}$ for 4 adults and 12 children
  y_i: the class: are they an adult/child?

(1) → Estimate prior: $P(a) = \frac{4}{4+12} = 0.25$, $P(c) = 0.75$

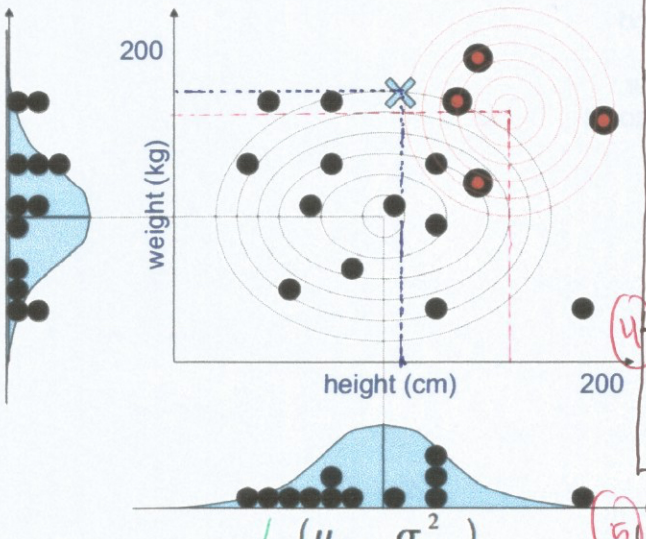
(2) → Assume heights and weights are modelled by gaussian distribution:

Model for adults {
- Height — Gaussian w/ mean, variance $\begin{cases} \mu_{h,a} = \frac{1}{4}\sum_{i:y_i=a} h_i \\ \sigma^2_{h,a} = \frac{1}{4}\sum_{i:y_i=a} (h_i - \mu_{h,a})^2 \end{cases}$
- Weight — Gaussian $(\mu_{w,a}, \sigma^2_{w,a})$
- Assume height & weight are independent.

↖ Maximum Likelihood Estimate (MLE) of the mean and variance

Model for children {
- Height $(\mu_{h,c}, \sigma^2_{h,c})$
- Weight $(\mu_{w,c}, \sigma^2_{w,c})$

$$P(a) = \frac{4}{4+12} = 0.25 \; ; \; P(c) = 0.75$$



$(\mu_{h,c}, \sigma^2_{h,c})$

Not a very good fit; but then again we don't have very many points to begin w/

This individual has a height that looks a lot like a child's height; it falls very close to the gaussian distr. for children's height

(3)

$$p(h_x|c) = \frac{1}{\sqrt{2\pi\sigma^2_{h,c}}} \exp -\frac{1}{2}\left(\frac{(h_x-\mu_{h,c})^2}{\sigma^2_{h,c}}\right)$$ = going to be a big number

$$p(w_x|c) = \frac{1}{\sqrt{2\pi\sigma^2_{w,c}}} \exp -\frac{1}{2}\left(\frac{(w_x-\mu_{w,c})^2}{\sigma^2_{w,c}}\right)$$ = going to be small number

↳ weight is a lot higher than the expected weight of a child

$$p(h_x|a) = \frac{1}{\sqrt{2\pi\sigma^2_{h,a}}} \exp -\frac{1}{2}\left(\frac{(h_x-\mu_{h,a})^2}{\sigma^2_{h,a}}\right)$$ = going to be a small number

$$p(w_x|a) = \frac{1}{\sqrt{2\pi\sigma^2_{w,a}}} \exp -\frac{1}{2}\left(\frac{(w_x-\mu_{w,a})^2}{\sigma^2_{w,a}}\right)$$ = going to be a big number

(4)

$$P(x|a) = p(h_x|a)\,p(w_x|a)$$

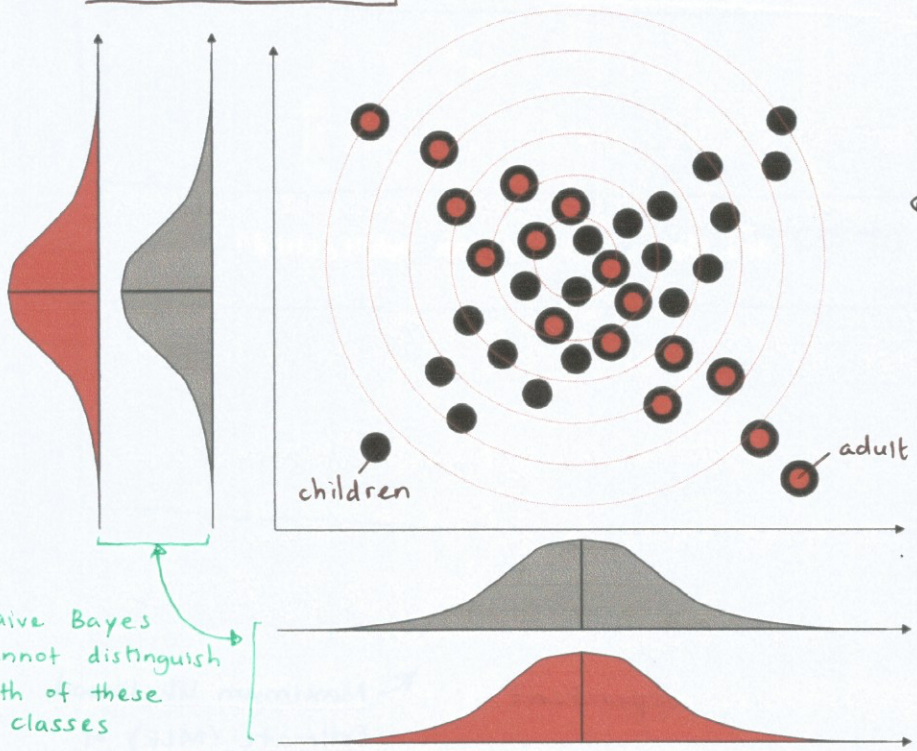
$$P(x|c) = p(h_x|c)\,p(w_x|c)$$

(5)

$$P(a|x) = \frac{P(x|a)P(a)}{P(x|a)P(a) + P(x|c)P(c)}$$

PROBLEMS W/ NAIVE BAYES

1) Correlation



- What differentiates the 2 classes is not the mean height/mean weight/varrance, but it is the co-variance.

~~Black class is (+)ively correlated~~

- There is (+)ve correlation for black, (-)ve correlation for red class; and that is the only difference

↓

Naive Bayes cannot do correlation!

2) Misclassification + Assuming word independence

- Separate spam from valid email, attributes = words

D1: "send us your password"	spam
D2: "send us your review"	ham
D3: "review your password"	ham
D4: "review us"	spam
D5: "send your password"	spam
D6: "send us your account"	spam

new email: "review us now"

P (spam) = 4/6 P (ham) = 2/6

spam	ham	
2/4	1/2	password
1/4	2/2	review
3/4	1/2	send
3/4	1/2	us
3/4	1/2	your
1/4	0/2	account

→ $P(\text{spam} \mid \text{"review us"}) \propto P(\text{spam})\ P(0,1,0,1,0,0 \mid \text{spam})$

$\propto 4/6 \cdot (1-2/4)(1/4)(1-3/4)(3/4)(1-3/4)(1-1/4)$

≈ 0.00293

→ $P(\text{ham} \mid \text{"review us"}) \propto 2/6 \cdot (1-1/2)(2/2)(1-1/2)(1/2)(1-1/2)(1-0/2)$

$\approx \boxed{0.02083} > 0.00293$

→ Classify the email as ham. However, notice the email is same as D4 which is a spam ~~message~~. training example.

↳ The only thing we did wrong is assume independence.

→ Naive Bayes allow this to happen.

3) Zero-frequency problem

$P(\text{'account'} \mid \text{ham}) = \frac{0}{2}$

→ The word 'account' has zero probability in ham emails

→ SOLUTION: Laplace Smoothing

- Add a small (+)ve no. to all counts
- $\varepsilon = \frac{\text{num}(w)}{\text{num}}$

MISSING VALUES IN NAIVE BAYES

→ Say we want to compute $P(X_1 = x_1, \ldots, X_J = ?, \ldots, X_d = x_d \mid y)$ but we don't know the value for some attribute X_J

- W/ Naive Bayes, you can simply ignore this attribute, hence

$$P(x_1 \ldots X_J \ldots x_d \mid y) = \prod_{i \neq J}^{d} P(x_i \mid y)$$

- This is based on conditional independence between attributes.