Naïve Bayes

~Abhishek Kumar

Scope

- What is Bayes Theorem?
- Components and their function
- Independence Assumption
- Continuous example
- Discrete example
- Pros and Cons
- Implementation in Python

Types of Machine learning

- **Supervised learning**
  - **Classification**
  - Regression
- Unsupervised learning
- Reinforcement learning

Bayesian statistic

- Bayesian statistics is a theory in the field of statistics based on the Bayesian interpretation of probability where probability expresses a degree of belief in an event.

- The degree of belief may be based on **prior** knowledge about the event.

Spam Detector

100 e-mails

Spam Detector

25 Spam

75 No spam

"Buy"

25 Spam

75 No spam

"Buy"

25 Spam

75 No spam

Spam Detector

 "Buy"

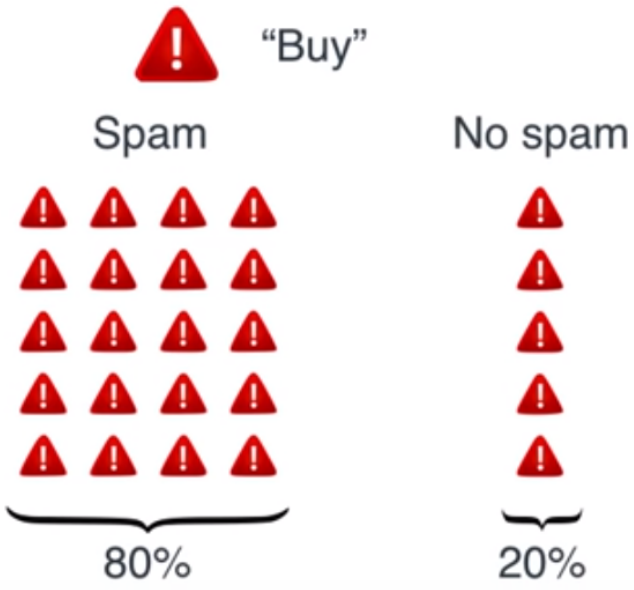
Spam

No spam

Quiz: If an e-mail contains the word "buy", what is the probability that it is spam?

- ○ 40%
- ○ 60%
- ○ 80%
- ○ 100%

Spam Detector



Quiz: If an e-mail contains the word "buy", what is the probability that it is spam?

- ○ 40%
- ○ 60%
- ● 80%
- ○ 100%

Solution:
80%

"Cheap"

Spam

No spam

"Cheap"

Spam

No spam

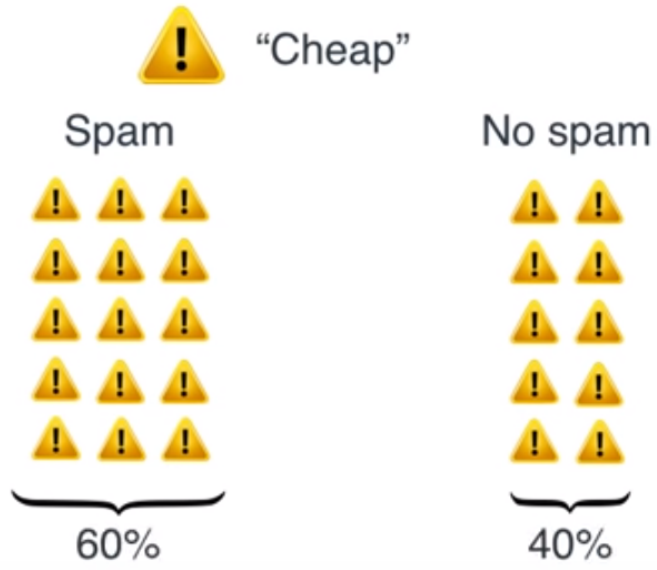
 "Cheap"

Spam

No spam

Quiz: If an e-mail contains the word "cheap", what is the probability that it is spam?

- ○ 40%
- ○ 60%
- ○ 80%
- ○ 100%



Quiz: If an e-mail contains the word “cheap”, what is the probability that it is spam?

- ○ 40%
- ● 60%
- ○ 80%
- ○ 100%

"Buy" and "Cheap"

Spam

No spam

"Buy" and "Cheap"

Spam

No spam

"Buy" and "Cheap"

Spam

No spam

"Buy" and "Cheap"

Spam

No spam

"Buy" and "Cheap"

Spam

No spam

"Buy" and "Cheap"

Spam

No spam

Quiz: If an e-mail contains the words "buy" and "cheap", what is the probability that it is spam?

- ○ 40%
- ○ 60%
- ○ 80%
- ○ 100%

"Buy" and "Cheap" ⟶ 100% ?

Spam

100%

No spam

0%

Quiz: If an e-mail contains the words "buy" and "cheap", what is the probability that it is spam?

- ○ 40%
- ○ 60%
- ○ 80%
- ● 100%

Solution:
100%

"Buy" and "Cheap"

Spam

12 e-mails

No spam

0 e-mails?


Solution: Collect more data?

Spam Detector

"Buy" and "Cheap"

Spam

12 e-mails

No spam

0 e-mails?

100 e-mails

100 e-mails

5 “Buy” 5% “Buy”

10 “Cheap” 10% “Cheap”

100 e-mails

5 “Buy”

10 “Cheap”

5% “Buy”

10% “Cheap”

0.5% “Buy” and “Cheap”

100 e-mails

5 “Buy”

10 “Cheap”

5% “Buy”

10% “Cheap”

Independent

0.5% “Buy” and “Cheap”


That's naive!
100 e-mails
5 "Buy"
10 "Cheap"
5% "Buy"
10% "Cheap"
Independent
0.5% "Buy" and "Cheap"

Spam Detector

Spam

25 e-mails
20 “Buy”
15 Cheap

4/5
3/5 → 12/25 × 25 = 12 “Buy” and “Cheap”

Spam Detector

No spam

75 e-mails

5 “Buy” 1/15

10 “Cheap” 2/15 → 2/225 × 75 = 2/3 “Buy” and “Cheap”

Spam Detector

"Buy" and "Cheap"

Spam

12

No spam

2/3

Spam Detector

"Buy" and "Cheap"

Spam

12

No spam

2/3

Quiz: If an e-mail contains the words "buy" and "cheap", what is the probability that it is spam?

Spam Detector

"Buy" and "Cheap"

Spam | No spam

12 | 2/3

94.737% | 5.263%

Quiz: If an e-mail contains the words "buy" and "cheap", what is the probability that it is spam?

$$\frac{12}{12 + 2/3} = \frac{36}{38}$$

$$= 94.737\%$$

Goal

- Learning function f(x) → y
- y = one of the classes (eg: Spam/ham , digits 0-9)
- x = x1.....xd – value of attributes (numerical or categorical)

Probabilistic Classifier

- Most probable class given observation

$$\hat{y} = \underset{y}{arg\,max}\, P(y|x)$$

- Bayesian Probability of class:

$$P(y|x) = \frac{\overbrace{P(x|y)}^{\text{class model}}\overbrace{P(y)}^{\text{prior}}}{\sum P(x|y')P(y')}$$

Bayesian Classification: Components

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

Example:
y … UK patient has Ebola
x … observed symptoms

P(y): Prior probability of each class

- Encodes which classes are rare or common
- Apriori much more likely to have something other than Ebola

Bayesian Classification: Components

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

Example:
y … UK patient has Ebola
x … observed symptoms

P(x/y): class conditional model

- Describe how likely to see an observation x for class y
- Assuming that its Ebola, do symptoms look plausible?

Bayesian Classification: Components

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

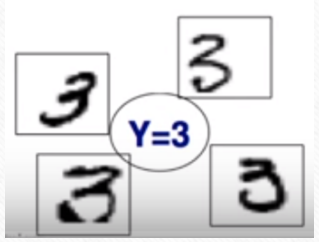
Example:
y … UK patient has Ebola
x … observed symptoms

P(x): normalize probabilities across observations

- Doesn't affect which class is most likely (argmax)

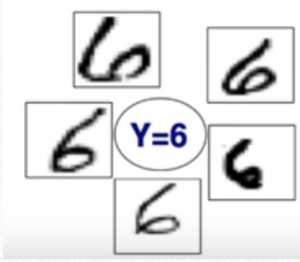
Why Normalization?

Normalizer: $P(x)=\sum_{y'} P(x|y')P(y')$



An outlier has a low probability in each class.

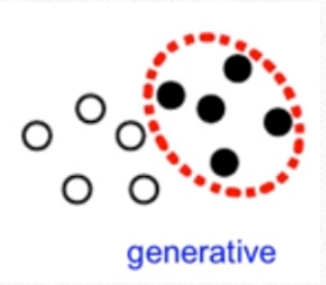
Normalizer makes comparable to non-outliers.

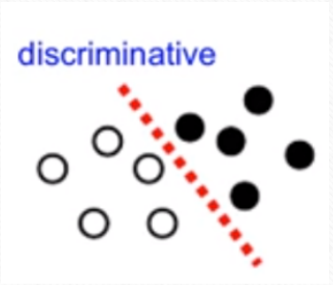


Generative model

- A complete probability distribution of each class

- All generative models are probabilistic classifiers but all probabilistic classifiers not necessarily generative model





Independence Assumption

- Compute P (x1….xd/y) for every observation x1….xd

- digits: 2^{400} possible black/white patterns (20x20)
- spam: every possible combination of words: $2^{10,000}$

- Idea: assume x1 ….. xd conditionally independent given y

Conditional Independence

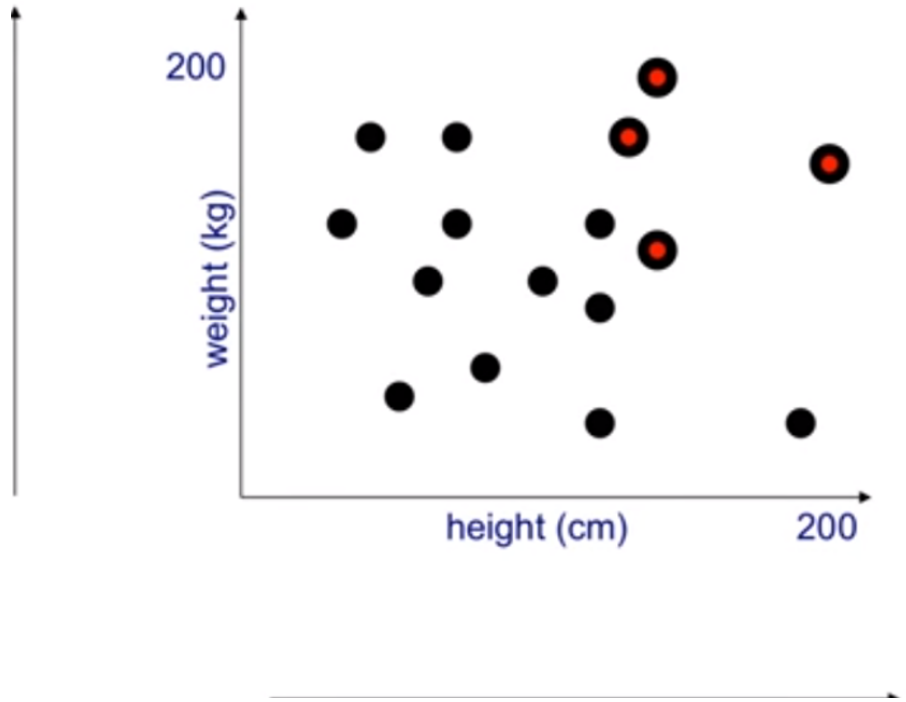
- Probabilities of going to beach and getting a heat stroke are not independent
- May be independent if we know weather is hot.
- Hot weather explains all the dependence between beach and heat stroke
- Class value explains all the dependence between attributes

Types of Naïve Bayes Models

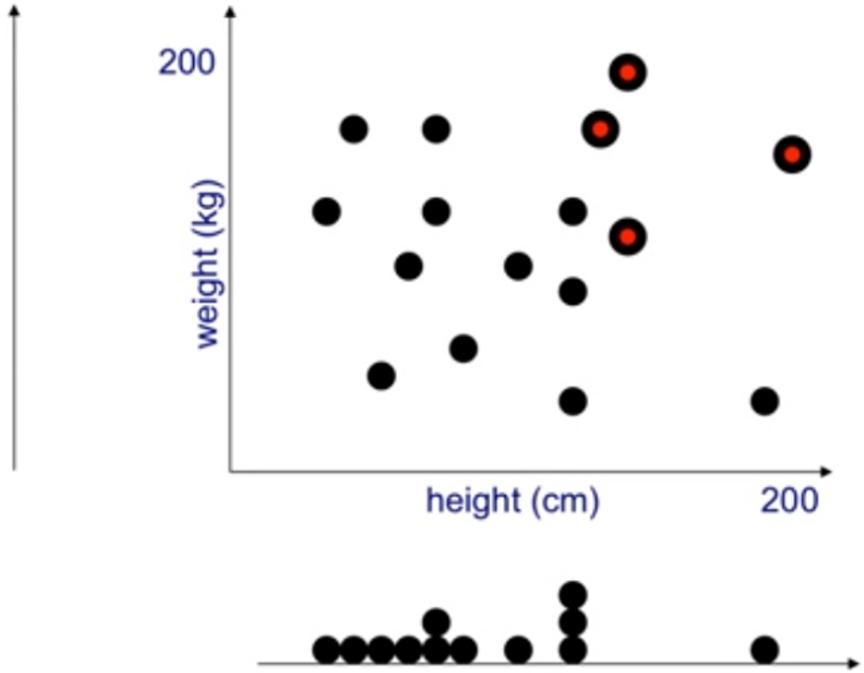
- **Multinomial**: good when your features describe discrete frequency counts(e.g: word counts)
- **Bernoulli**: good for making prediction from binary features
- **Gaussian**: good for making prediction from normally distributed features

Continuous example

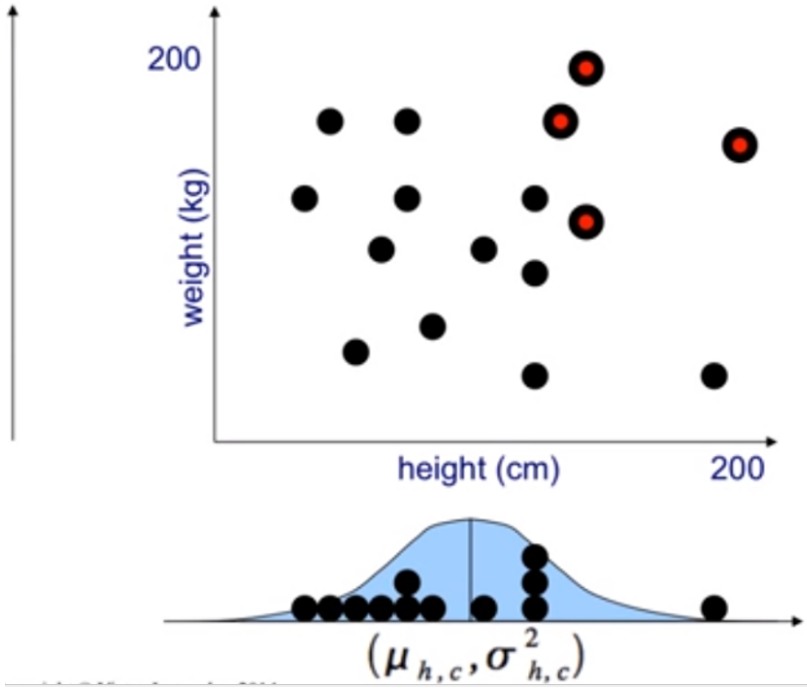
- Distinguish children from adults based on size
- classes: {adults ‘a’, children ‘c’} attributes: heights[cm] , weight[lb]
- training examples are 4 adults and 12 children
- Class probability P(a) 4/4+12 = 0.25 , P(c) = 0.75
- Model height and weight : Gaussian with mean with variance
- Assuming height and weight are independent


200
weight (kg)
height (cm)
200

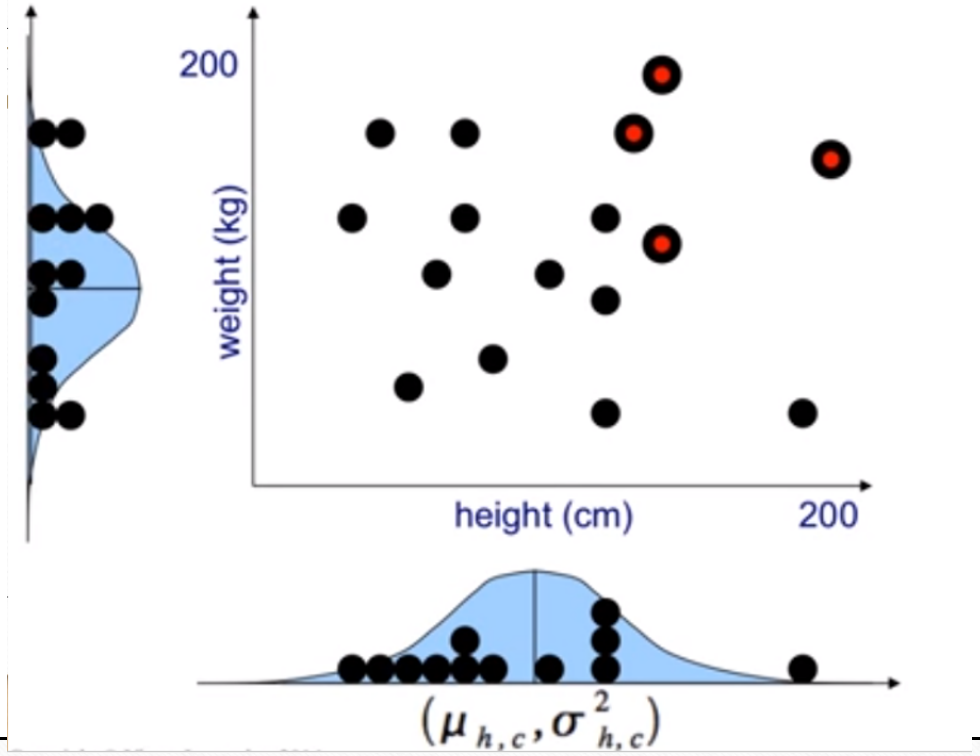
$$P(a)=\frac{4}{4+12}=0.25\ ;\ P(c)=0.75$$



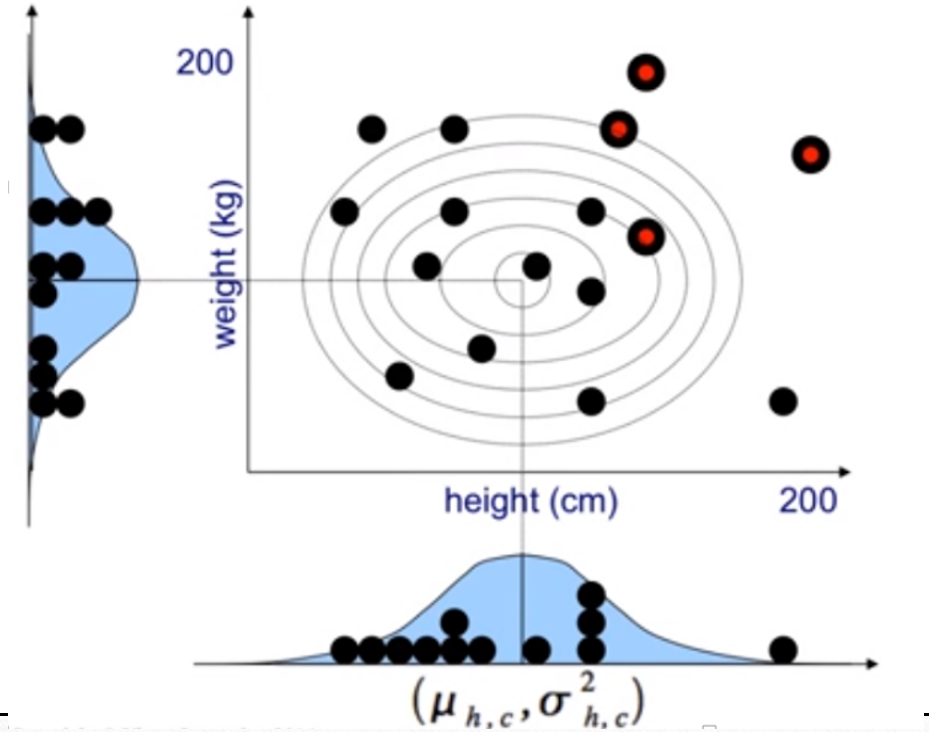
$$P(a)=\frac{4}{4+12}=0.25\ ;\ P(c)=0.75$$



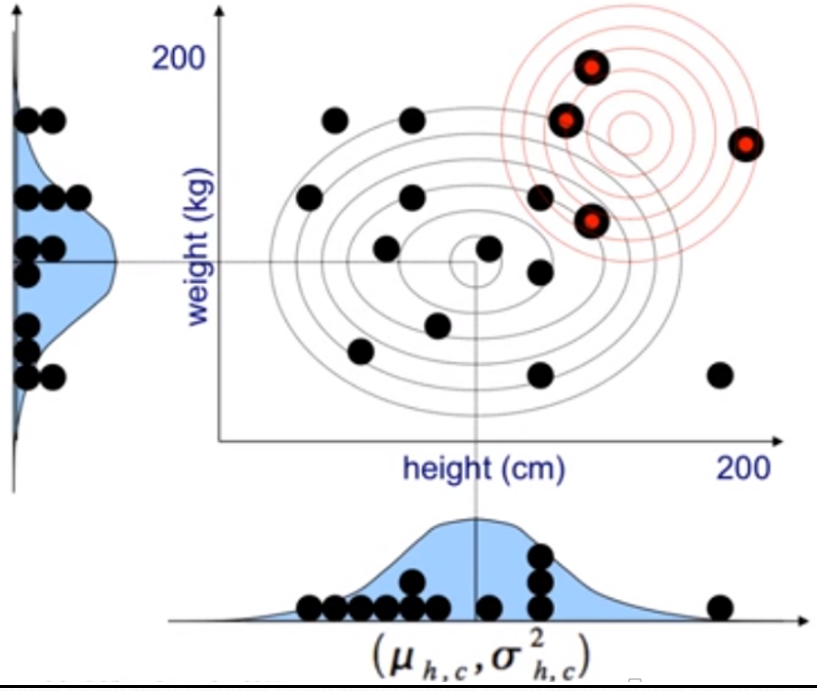
$$P(a)=\frac{4}{4+12}=0.25\ ;\ P(c)=0.75$$



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

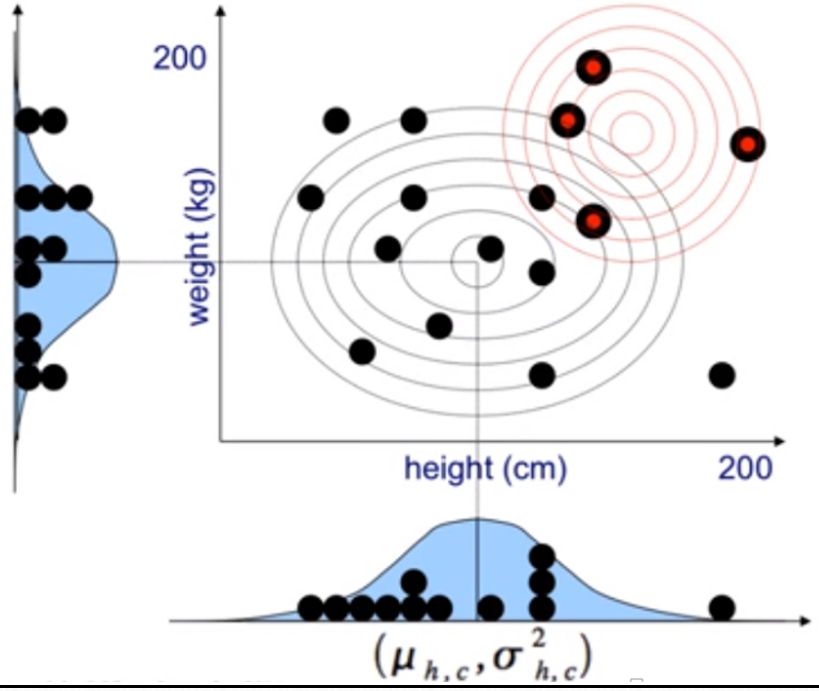


$$P(a)=\frac{4}{4+12}=0.25\,;\,P(c)=0.75$$

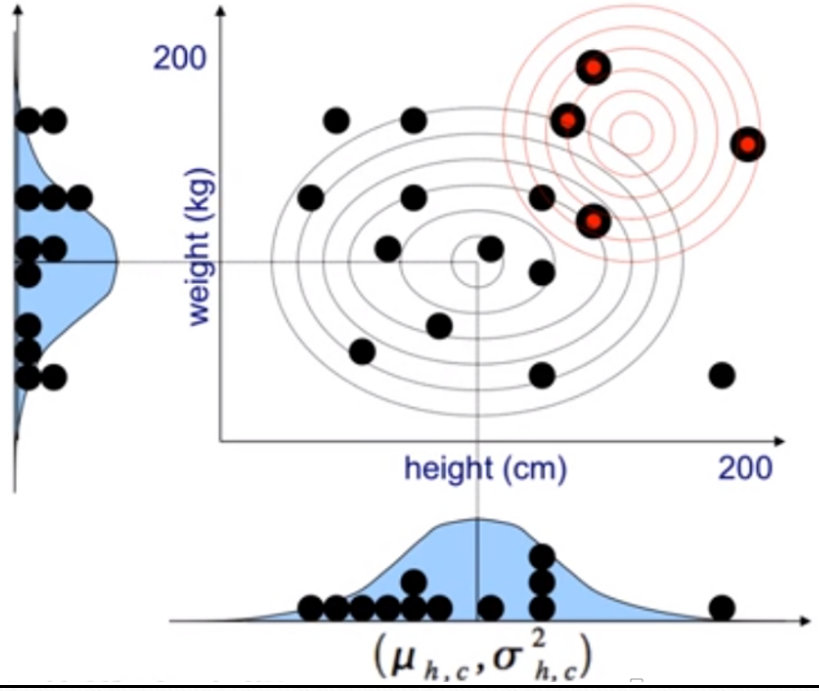


$$P(a)=\frac{4}{4+12}=0.25\ ;\ P(c)=0.75$$

$$p(h_x|c)=\frac{1}{\sqrt{2\pi\,\sigma_{h,c}^2}}\exp-\frac{1}{2}\left(\frac{(h_x-\mu_{h,c})^2}{\sigma_{h,c}^2}\right)$$



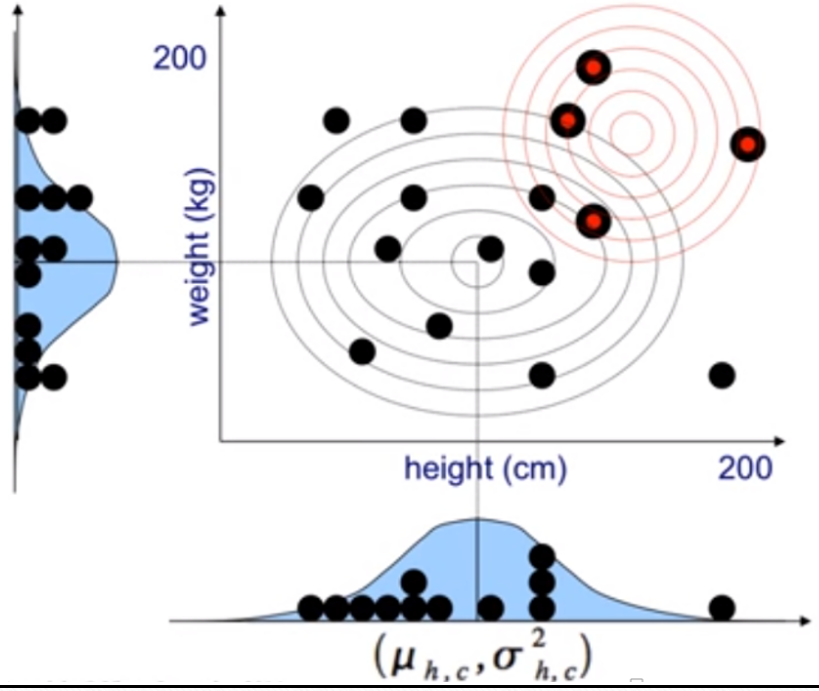
$$P(a)=\frac{4}{4+12}=0.25\ ;\ P(c)=0.75$$



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

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

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



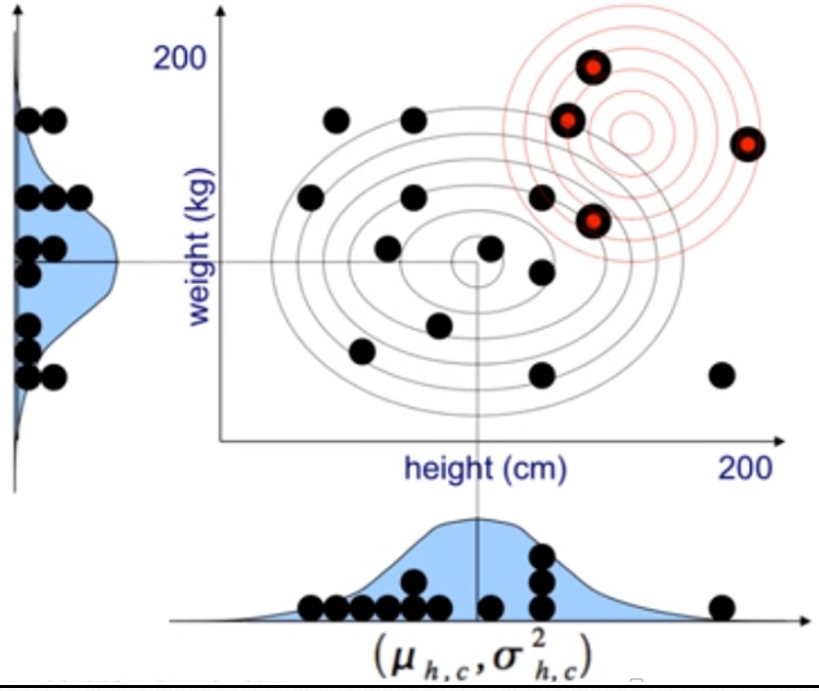
$$p(h_x|c)=\frac{1}{\sqrt{2\pi\,\sigma^2_{h,c}}}\exp-\frac{1}{2}\left(\frac{(h_x-\mu_{h,c})^2}{\sigma^2_{h,c}}\right)$$

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

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

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

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

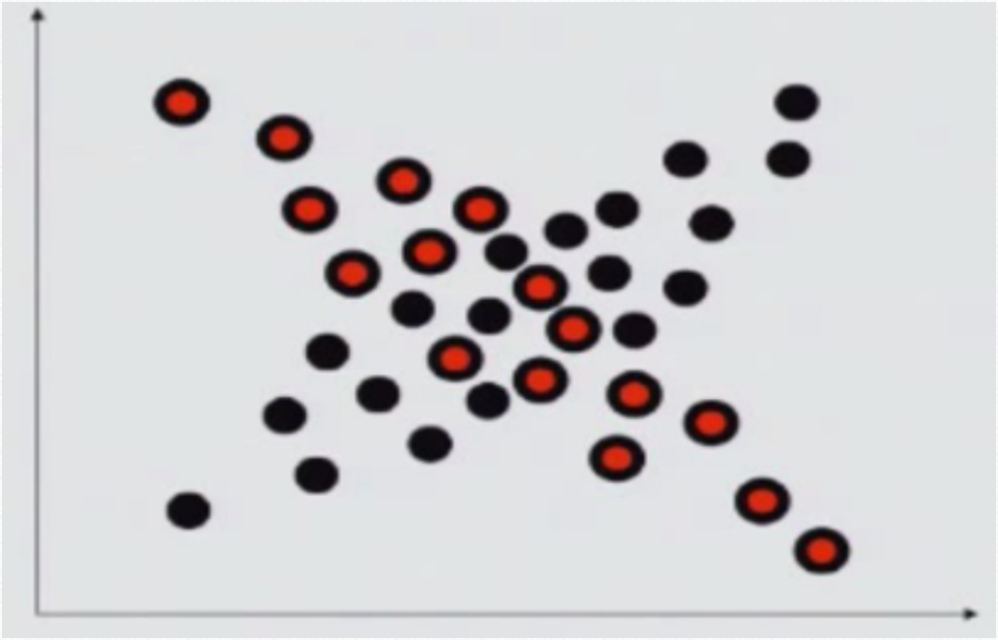


$$P(x|a)=p(h_x|a)\,p(w_x|a)$$

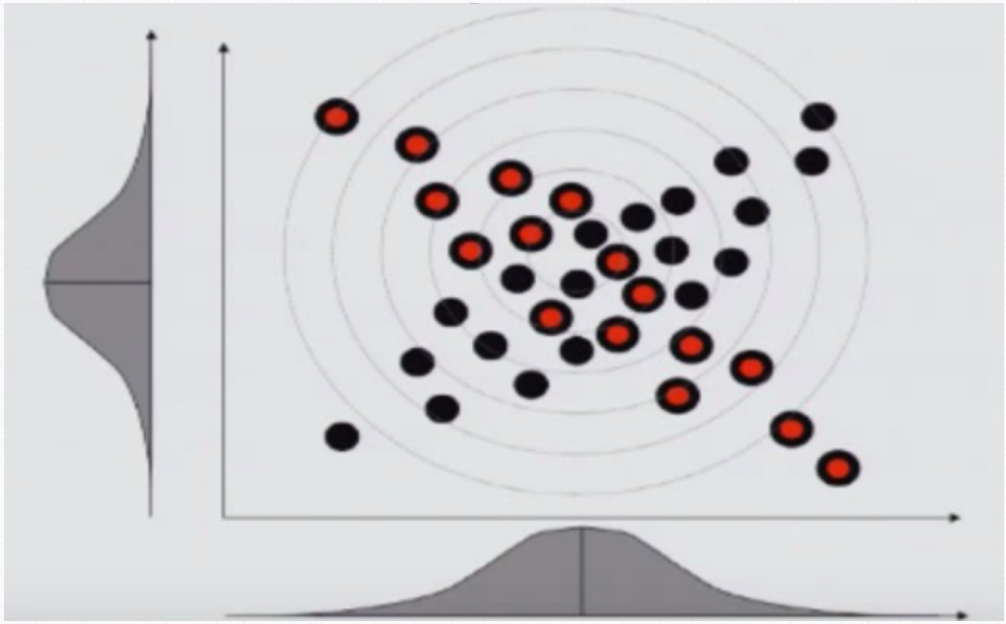
$$P(x|c)=p(h_x|c)\,p(w_x|c)$$

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

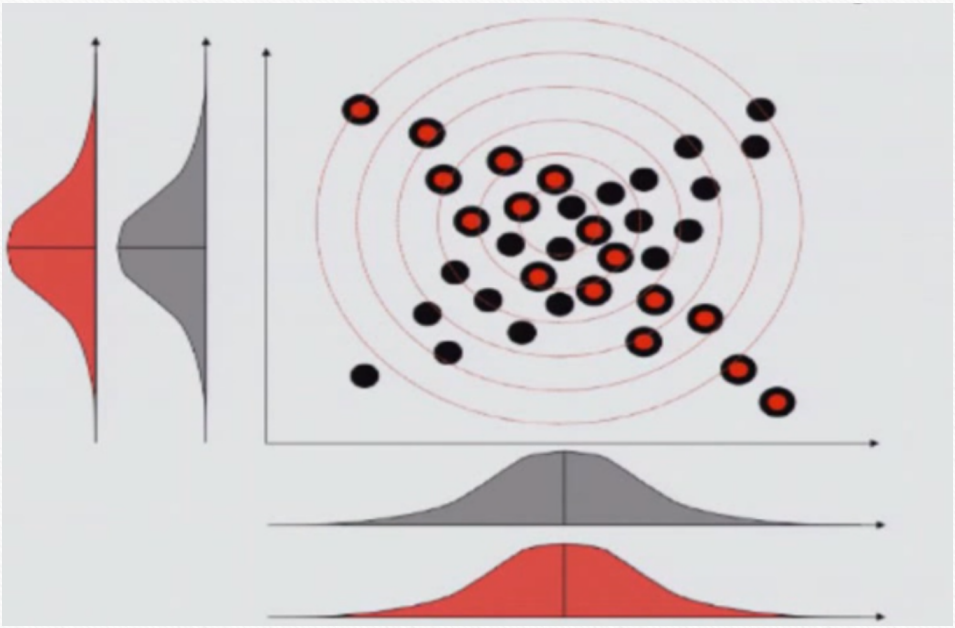
Problem With Naïve Bayes



Problem With Naïve Bayes



Problem With Naïve Bayes



Discrete Example: Spam

- Separate words from valid email, attributes are words
- D1: “send us your password” spam
- D2: “send us your reviews” ham
- D3: “review your password” ham
- D4: “review us” spam
- D5: “send your password” spam
- D6: “send us your account” spam

Discrete Example: Spam

- Separate words from valid email, attributes are words
- D1: "send us your password" spam
- D2: "send us your reviews" ham
- D3: "review your password" ham
- D4: "review us" spam
- D5: "send your password" spam
- D6: "send us your account" spam

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

Discrete Example: Spam

- Separate words from valid email, attributes are words

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

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

Discrete Example: Spam

- Separate words from valid email, attributes are words
- D1: "send us your password" spam
- D2: "send us your reviews" ham
- D3: "review your password" ham
- D4: "review us" spam
- D5: "send your password" spam
- D6: "send us your account" spam

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

- New email: "review us now"

Discrete Example: Spam

- Separate words from valid email, attributes are words
- D1: "send us your password" spam
- D2: "send us your reviews" ham
- D3: "review your password" ham
- D4: "review us" spam
- D5: "send your password" spam

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

$$P(\text{review us}\,|\text{spam}) = P(0,1,0,1,0,0\,|\text{spam}) = (1-\frac{2}{4})(\frac{1}{4})(1-\frac{3}{4})(\frac{3}{4})(1-\frac{3}{4})(1-\frac{1}{4})$$

$$P(\text{review us}\,|\text{ham}) = P(0,1,0,1,0,0\,|\text{ham}) = (1-\frac{1}{2})(\frac{2}{2})(1-\frac{1}{2})(\frac{1}{2})(1-\frac{1}{2})(1-\frac{0}{2})$$

Discrete Example: Spam

- Separate words from valid email, attributes are words

- D1: "send us your password" spam
- D2: "send us your reviews" 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(ham | review us) = 0.87

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

Problem With Naïve Bayes

- Zero frequency problem:

  any email containing 'account' is spam P(account|ham) = 0

  Solution: never allows zero probabilities

  Laplace smoothing: Add a small positive number to all counts

  It's a very common problem (Zipf's law: 50% words occur once)

Problem With Naïve Bayes

- Assume word independence:

  Every word contributes independently to P(spam|email)

  Fooling Naïve Bayes by adding lots of hammy words into spam email.

Advantages

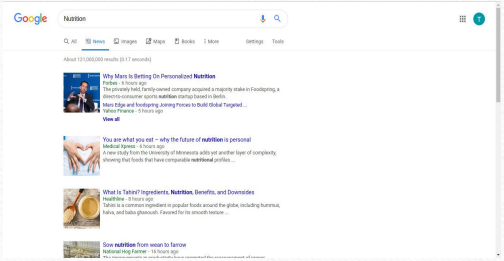
1. Very simple and easy to implement
2. Need less training data
3. Handle both continuous and discrete data
4. As it is fast, it can be used in real time prediction
5. Not sensitive to irrelevant features
6. No requirement of fill in or explicitly model missing value

Applications in real world

Facial recognition

Weather Prediction

Medical Diagnosis



News Classification

Python code implementation

- Jupyter Notebook

Discussion

Thank you!!!!!!