Machine Learning – COMS3007

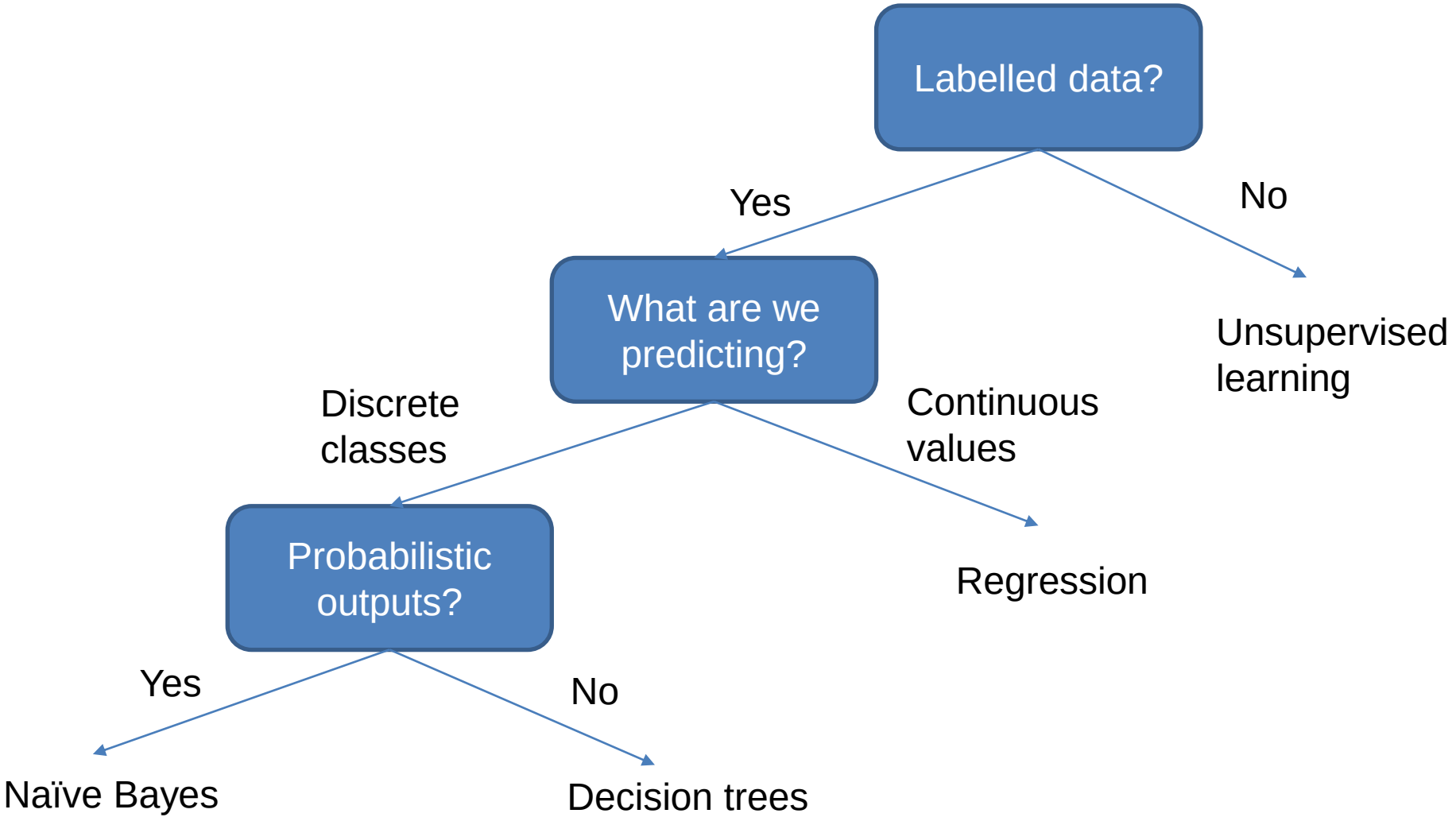
More on Naïve Bayes

Benjamin Rosman

Based heavily on course notes by Chris Williams and Victor Lavrenko, Chris Thornton, and Clint van Alten

More info in:
Machine Learning by Tom Mitchell (1997), Section 6.9

When to use Naïve Bayes?



Example Question

- A builder is sneezing. What is wrong with him?

- Features:
  - Symptom (sneezing)
  - Occupation (builder)

- Class:
  - Ailment (hayfever, flu)

Training data

Symptom	Occupation	Ailment
Sneezing	Nurse	Flu
Sneezing	Farmer	Hayfever
Headache	Builder	Hayfever
Headache	Builder	Flu
Sneezing	Teacher	Flu
Sneezing	Teacher	Flu

- We now want to be able to query:
  - $P(flu \mid sneezing, builder)$
  - $P(hayfever \mid sneezing, builder)$
- Choose the maximum (MAP solution)

Priors

Symptom	Occupation	Ailment
Sneezing	Nurse	Flu
Sneezing	Farmer	Hayfever
Headache	Builder	Hayfever
Headache	Builder	Flu
Sneezing	Teacher	Flu
Sneezing	Teacher	Flu

- What is the prior probability of having flu?
  - $P(flu) = 4\ (cases\ of\ flu)\ /\ 6\ (data\ points) = \frac{2}{3}$
  - $P(hayfever) = \frac{2}{6} = \frac{1}{3}$

Consider flu

Symptom	Occupation	Ailment
Sneezing	Nurse	Flu
Sneezing	Farmer	Hayfever
Headache	Builder	Hayfever
Headache	Builder	Flu
Sneezing	Teacher	Flu
Sneezing	Teacher	Flu

We want $P(flu \mid sneezing, builder)$

Recall Bayes' Rule:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

So:

$$p(flu|sneezing, builder) = \frac{p(sneezing, builder|flu)p(flu)}{p(sneezing, builder)}$$

How do we deal with $p(sneezing, builder|flu)$?

Naïve Bayes assumption:

$$p(sneezing, builder|flu) = p(sneezing|flu)\ p(builder|flu)$$

Class conditionals

$$p(sneezing, builder|flu) = p(sneezing|flu)\ p(builder|flu)$$

$$p(sneezing|flu) = \frac{number\ of\ sneezing\ AND\ flu}{number\ of\ flu} = \frac{3}{4}$$

$$p(builder|flu) = \frac{1}{4}$$

Symptom	Occupation	Ailment
Sneezing	Nurse	Flu
Sneezing	Farmer	Hayfever
Headache	Builder	Hayfever
Headache	Builder	Flu
Sneezing	Teacher	Flu
Sneezing	Teacher	Flu

$$p(sneezing|hayfever) = \frac{1}{2}$$

$$p(builder|hayfever) = \frac{1}{2}$$

So

$$p(sneezing, builder|flu) = \frac{3}{4} \times \frac{1}{4} = \frac{3}{16}$$

$$p(sneezing, builder|hayfever) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Class conditionals

$$p(sneezing, builder|flu) = \frac{3}{4} \times \frac{1}{4} = \frac{3}{16}$$

$$p(sneezing, builder|hayfever) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Why can't we say the builder has hayfever? (1/4 > 3/16)

The priors: $p(flu) = 2/3$ and $p(hayfever) = 1/3$

Flu is twice as likely as hayfever!

Symptom	Occupation	Ailment
Sneezing	Nurse	Flu
Sneezing	Farmer	Hayfever
Headache	Builder	Hayfever
Headache	Builder	Flu
Sneezing	Teacher	Flu
Sneezing	Teacher	Flu

Probabilities

We sum over the classes (this gives the denominator in Bayes' Rule and lets our probabilities sum to 1)

From sum rule:

$$p(sneezing, builder)$$
$$= p(sneezing, builder|flu)p(flu) + p(sneezing, builder|hayfever)p(hayfever)$$
$$= \left(\frac{3}{16} \times \frac{2}{3}\right) + \left(\frac{1}{4} \times \frac{1}{3}\right) = \frac{5}{24}$$

$$p(flu|sneezing, builder) = \frac{p(sneezing, builder|flu)p(flu)}{p(sneezing, builder)}$$

$$= \frac{\frac{3}{16} \times \frac{2}{3}}{\frac{5}{24}} = 0.6$$

$$p(hayfever|sneezing, builder) = \frac{p(sneezing, builder|hayfever)p(hayfever)}{p(sneezing, builder)}$$

$$= \frac{\frac{1}{4} \times \frac{1}{3}}{\frac{5}{24}} = 0.4$$

$$y^* = \arg\max_y P(y|x) = \arg\max\{0.6, 0.4\} = flu$$

A note about representations

- In the text examples, we had vectors like (1,0,1,0,0)
- Some important things:
  1. "1" means "this word EXISTS in this text", not "~~there is one word in the text~~": it is binary!
  2. Why do we multiply probabilities by (1-x) when there is a "0"?
     - If we didn't, we're just ignoring this feature!
     - The absence of a feature can be informative
       - E.g. The fact that I don't describe food as "terrible" or "disgusting" tells me as much as if I call it "good"
  3. Exact same representation as pixels (0/1)

Smoothing

- A nurse has a headache. What is the most likely ailment?

- Features:
  - Symptom (headache)
  - Occupation (nurse)

- Class:
  - Ailment (hayfever, flu)

Same data

Symptom	Occupation	Ailment
Sneezing	Nurse	Flu
Sneezing	Farmer	Hayfever
Headache	Builder	Hayfever
Headache	Builder	Flu
Sneezing	Teacher	Flu
Sneezing	Teacher	Flu

- What is the prior probability of having flu?
  - $P(flu) = 4\,(cases\ of\ flu)\ /\ 6\,(data\ points) = \frac{2}{3}$
  - $P(hayfever) = \frac{2}{6} = \frac{1}{3}$
- We now want $P(hayfever \mid headache, nurse)$

Consider hayfever (hf)

$$p(headache, nurse|hf) = p(headache|hf)\ p(nurse|hf)$$

$$p(headache|hf) = \frac{number\ of\ headache\ AND\ hf}{number\ of\ hf} = \frac{1}{2}$$

$$p(nurse|hf) = \frac{0}{2}$$

This is no good! It means nurses cannot have hayfever!
In reality, it just didn't appear in our training set!

Symptom	Occupation	Ailment
Sneezing	Nurse	Flu
Sneezing	Farmer	Hayfever
Headache	Builder	Hayfever
Headache	Builder	Flu
Sneezing	Teacher	Flu
Sneezing	Teacher	Flu

Smoothing

$$p(nurse|hf) = \frac{number\ of\ nurse\ AND\ hf}{number\ of\ hf} = \frac{0}{2}$$

Smoothing:
Add an "imaginary" occurrence of (nurse, hayfever) to the data
Then the 0 will not happen.
As we see many more instances of real data, they will greatly outweigh the "imaginary" data – but will never be 0.
But:
This adds bias to nurses – we're making them a bit more likely but not changing any other!
So we add an "imaginary" datapoint per type of other occupation: {farmer, builder, teacher} (a total of 4)

$$p(nurse|hf) = \frac{0+1}{2+4} = \frac{1}{6}$$

Underflow

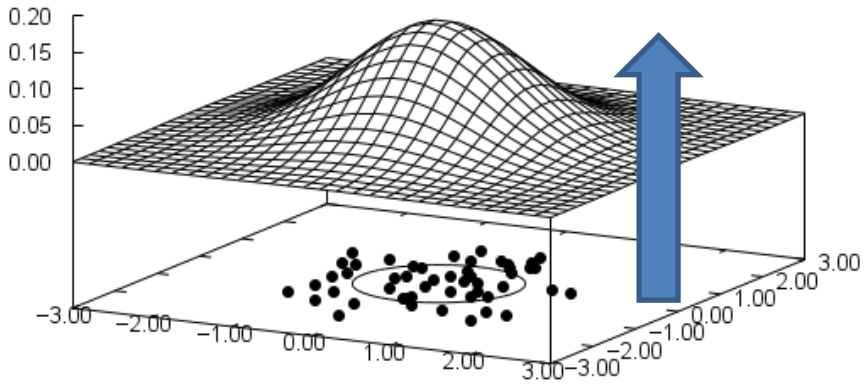
Looking at the questions in the lab, you may have found an underflow problem:

Multiplying many tiny values together can give such a small probability that it is rounded to 0.
e.g. $0.1 \times 0.2 \times 0.05 \times \cdots = 0.000000001 \approx 0$

Solution:
Use logs
Why?
$\log_{10} 0.000000001 = -9$
Multiplication and division → addition and subtraction
e.g. $\log\left(\frac{A*B}{C}\right) = \log(A) + \log(B) - \log(C)$

What to do with data?

- Estimate the distribution p from the data x
  - Learning problem: learn p that fits x
  - How to measure goodness of fit?
    - Given a distribution, how likely is it to have generated the data?
    - i.e. what is the probability of this data set given the distribution?



Recall: Likelihood

- Want the probability of data D given distribution/model M = likelihood of M = p(D|M)
  - $p(D|M) = \prod_{i=1}^{N} p(x_i|M)$
  - Product of probabilities of generating each data point independently
- Compute p(D|M) for different models M
- Pick M which gives the highest likelihood
  - This is the **maximum likelihood estimate**
  - Optimisation problem!

Example distribution: Gaussian

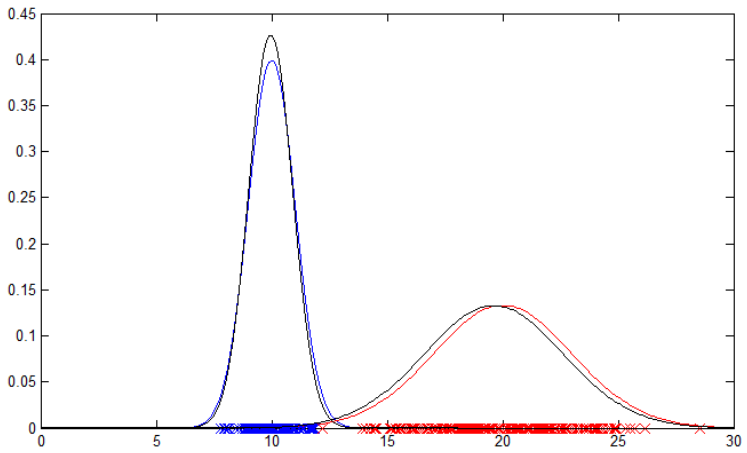
- Most common continuous distribution
- Often a very reasonable model
- Also called Normal distribution
- 1D Gaussian:

Normalisation factor: integrate to 1

mean

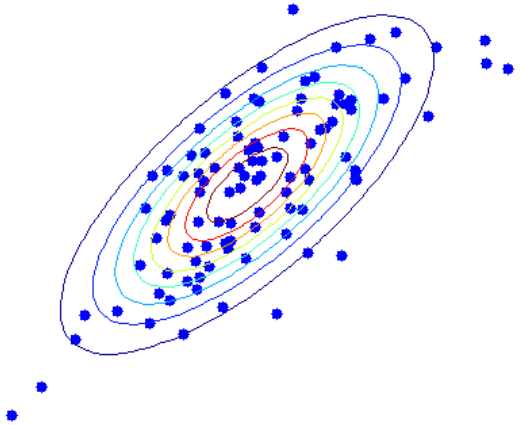
variance

$$p(x|\mu,\sigma^2) = N(x;\mu,\sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

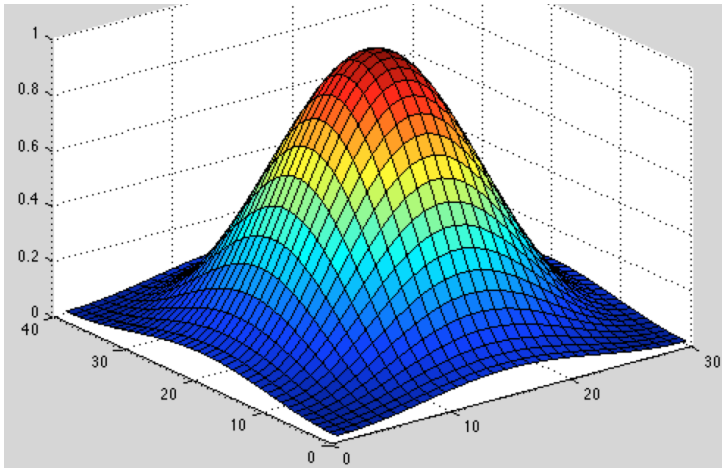


Probability of x decreases as x moves further from μ, with speed governed by σ

Multivariate Gaussian



- Data **x** is d-dimensional
- $p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$
- **μ** = d-dimensional mean vector
- Σ = covariance matrix: symmetric and positive definite
  - $\Sigma = E[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T]$
- Parameters to learn:
  - Σ = d(d+1)/2, **μ** = d. Why?



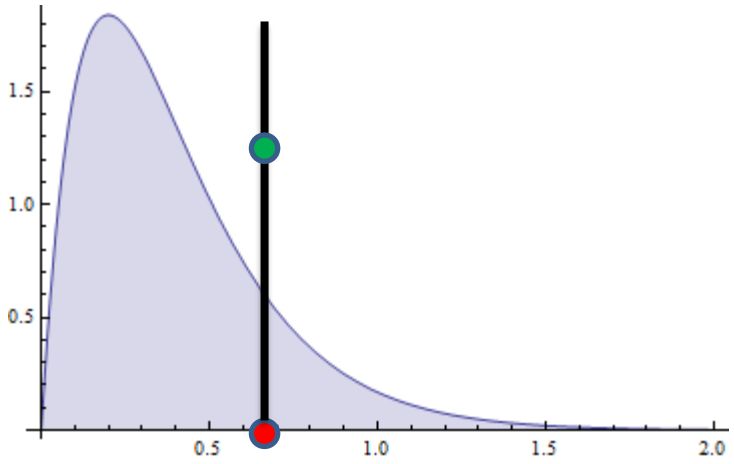
Max Likelihood Estimate of a Gaussian

- Given data $\{x_i, i = 1, 2, \ldots, n\}$
- MLE of the data given a 1D Gaussian model, gives:
  - $\hat{\mu} = \frac{\sum_i x_i}{n}$, $\hat{\sigma}^2 = \frac{\sum_i (x_i - \mu)^2}{n}$
  - How?
    - Compute prob of data given model: P(D|M)
    - We often compute the log probability instead
    - Differentiate and set = 0. Why?
    - Try this at home!
- MLE of multivariate Gaussian:
  - $\boldsymbol{\mu} = \frac{1}{n}\sum_i \boldsymbol{x}_i$
  - $\Sigma = \frac{1}{n}\sum_i (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$

Note: we do direct optimisation here: we only need to compute the MLE

Sampling

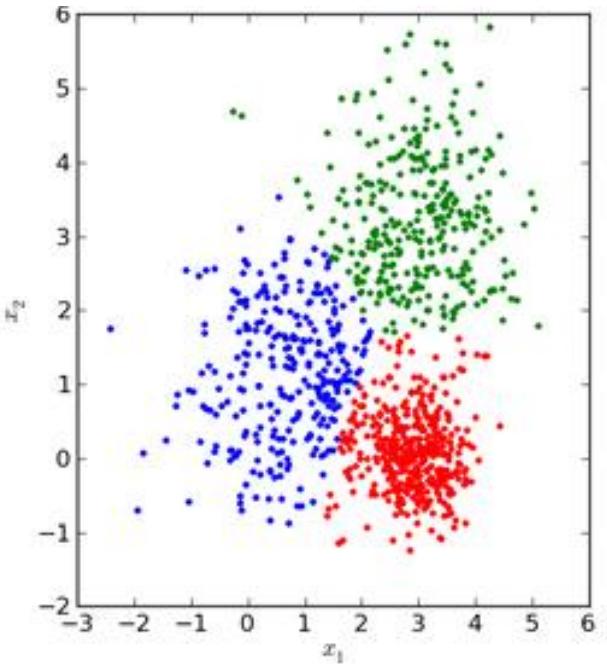
- Generate data x from distribution p (sampling)
  - How:
    - Pick a random point x in domain of p
    - Accept if random number < p(x)



Generating More Structured Data

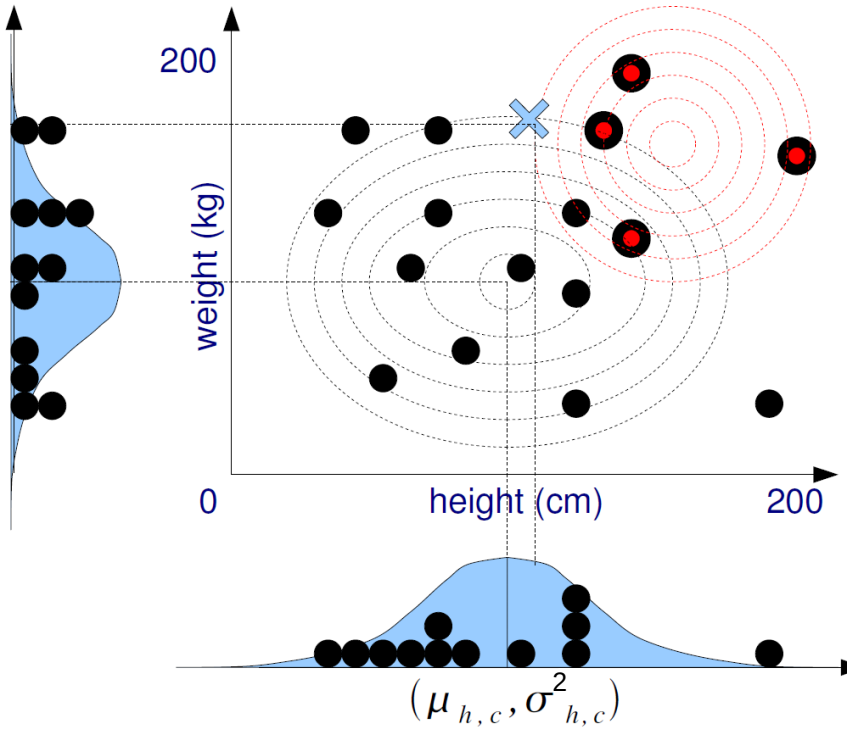
- Assume we want 2D data from 3 classes = {A, B, C}
- Define 3 Gaussians in 2D, label them A, B, C
  - Define a μ and Σ for each
- Define a prior for each (how likely is this class)
  - P(A) + P(B) + P(C) = 1

- To generate a data point:
  - Randomly pick a class y using the prior
  - Randomly sample a point **x** from model y
  - Add (**x**, y) to dataset

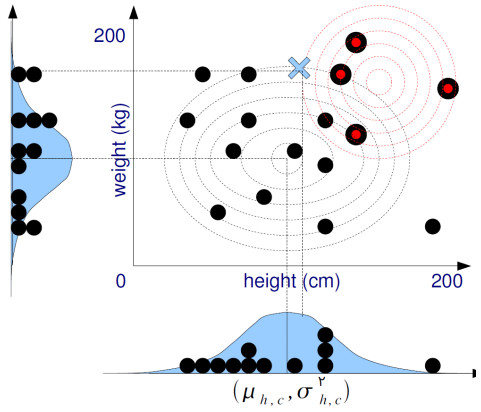


Continuous Example

- Distinguish children from adults based on size
  - Classes: {a, c}
  - Attributes: height (cm), weight (kg)
  - Training data:
    - $\{h_i, w_i, y_i\}$
  - 4 adults, 12 children



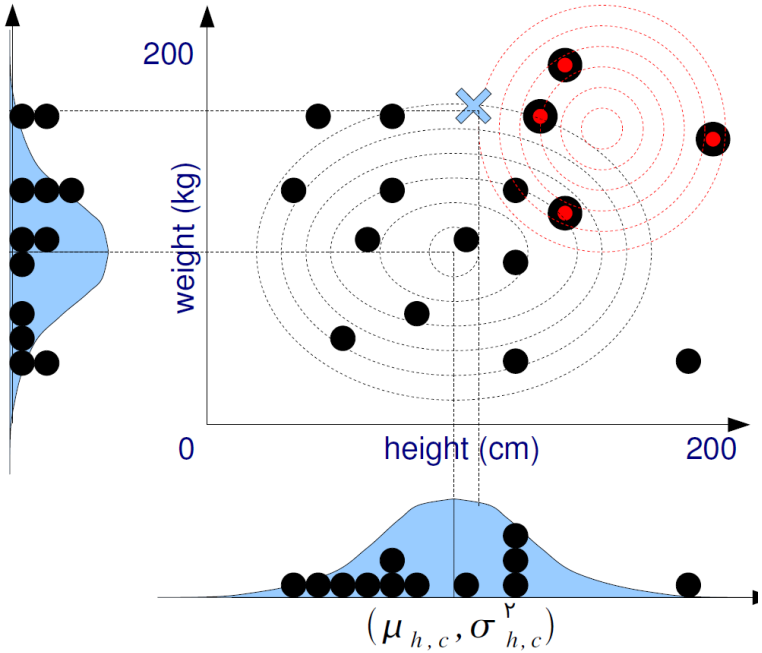
Continuous Example



- Class priors?
  - $P(a) = \frac{4}{4+12} = 0.25, P(c) = 0.75$
- Model for adults (assume independence)?
  - $height \sim N(\mu_{h,a}, \sigma^2_{h,a})$
    - $\mu_{h,a} = \frac{1}{4}\sum_{i:y_i=a} h_i$
    - $\sigma^2_{h,a} = \frac{1}{4}\sum_{i:y_i=a}(h_i - \mu_{h,a})^2$
  - $weight \sim N(\mu_{w,a}, \sigma^2_{w,a})$
- Similarly for children
  - $height \sim N(\mu_{h,c}, \sigma^2_{h,c}), weight \sim N(\mu_{w,c}, \sigma^2_{w,c})$

We have assumed adult heights are Gaussian.
So, work out the mean and variance of the heights of all the adults. Then you have fit a Gaussian to the data.

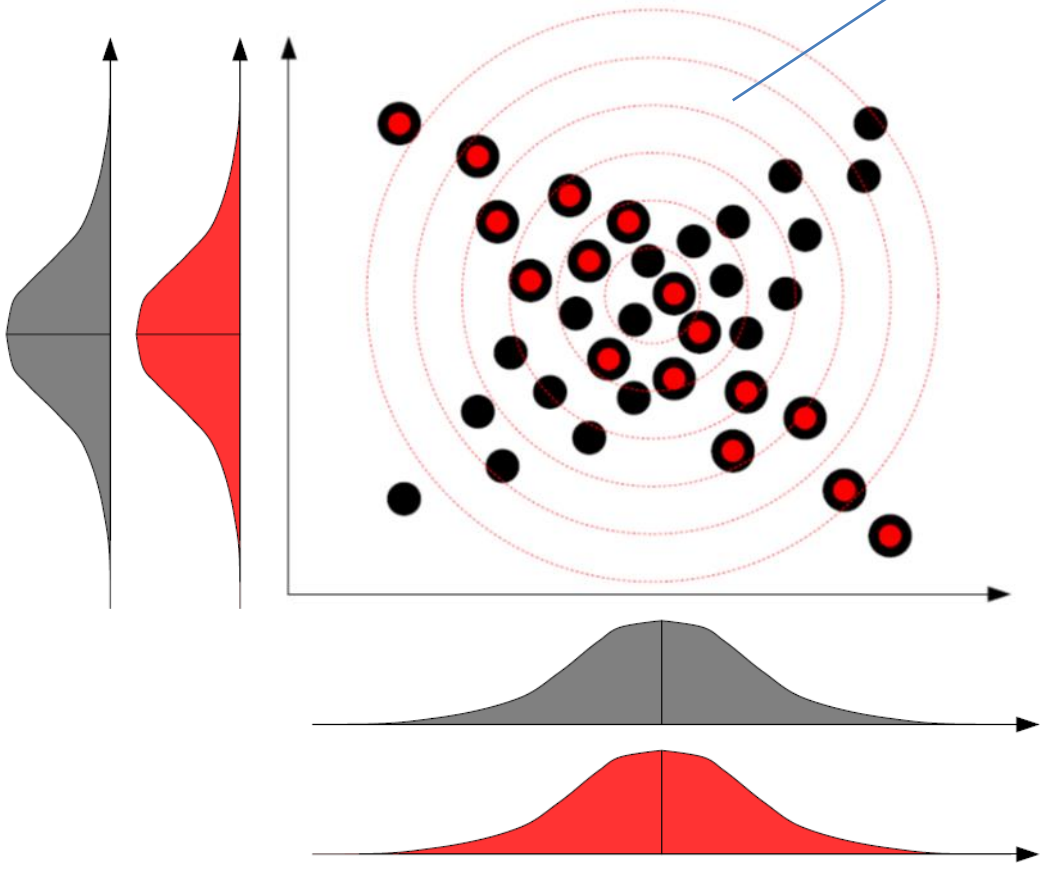
Example



- $P(a) = 0.25, P(c) = 0.75$
- $p(h_x|c) = \frac{1}{\sqrt{2\pi\sigma_{h,c}^2}} \exp\{-\frac{1}{2}(\frac{(h_x-\mu_{h,c})^2}{\sigma_{h,c}^2})\}$
- $p(w_x|c) = \frac{1}{\sqrt{2\pi\sigma_{w,c}^2}} \exp\{-\frac{1}{2}(\frac{(w_x-\mu_{w,c})^2}{\sigma_{w,c}^2})\}$
- Same for $p(h_x|a)$ and $p(w_x|a)$
- $p(x|c) = p(h_x|c)p(w_x|c)$
- $p(x|a) = p(h_x|a)p(w_x|a)$
- $p(c|x) = \frac{p(x|c)p(c)}{p(x|c)p(c)+p(x|a)p(a)}$

$$P(c|x) = \frac{\prod_{i=1}^{n} P(x_i|c)\, P(c)}{P(x)}$$

Being too Naïve



Easy to classify based on joint distribution $P(x_1, x_2|c)$

Impossible to classify based on marginal distributions $P(x_1|c)$ or $P(x_2|c)$:
Independence assumptions do not hold!

(Probabilistic) Graphical Models

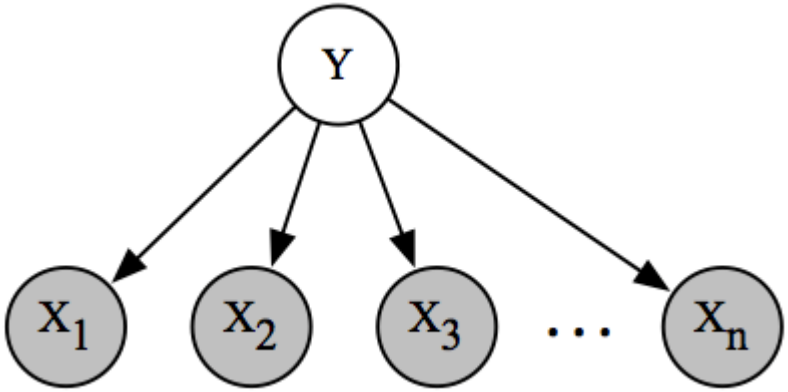
More general models…

Relationship between X and Y

X depends on Y

$$P(X,Y) = P(X|Y)P(Y)$$

Probabilistic Graphical Models (PGMs)

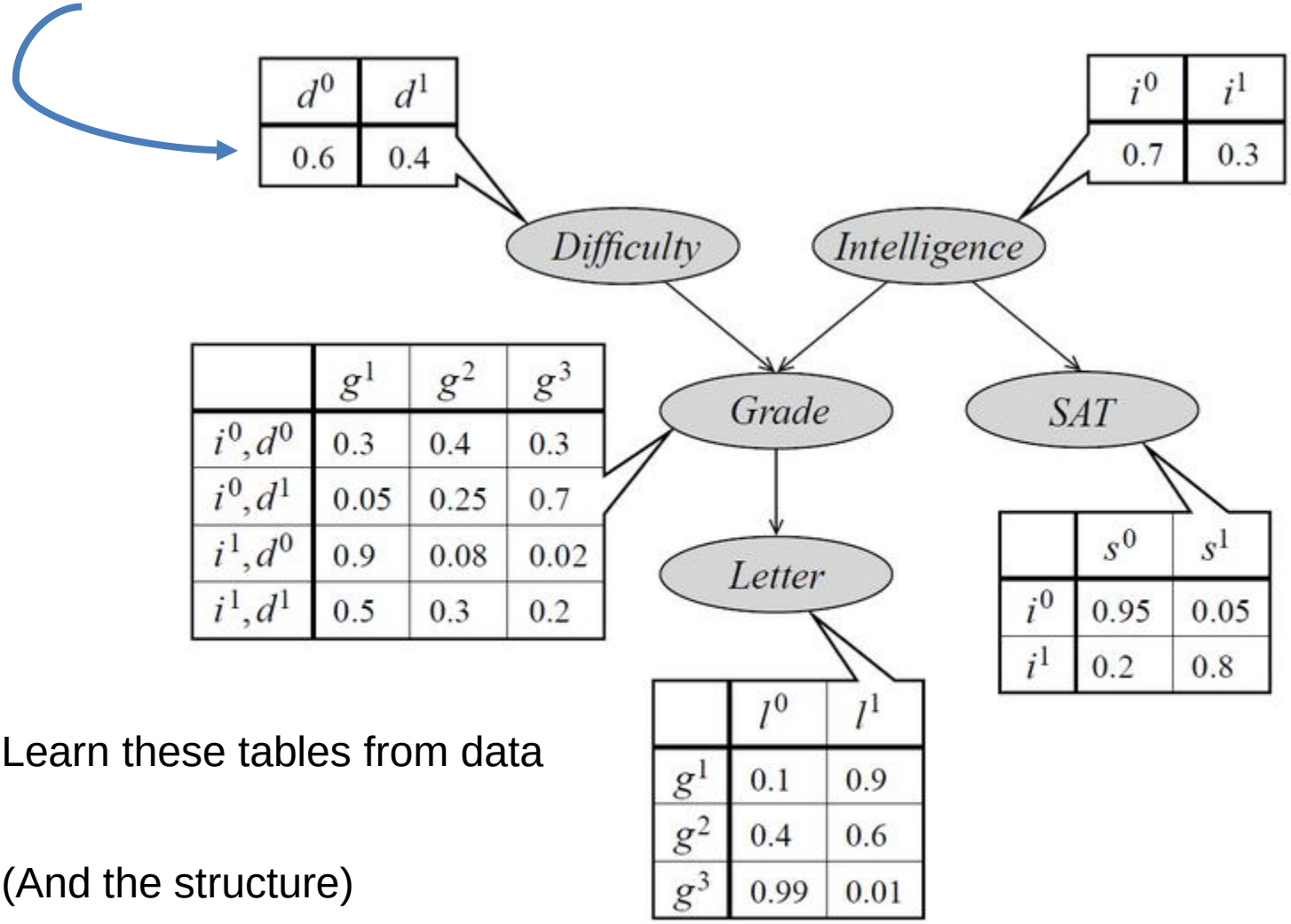


$$P(X_1, X_2, X_3, \ldots, X_n, Y) = P(X_1|Y)P(X_2|Y)P(X_3|Y) \ldots P(X_n|Y)P(Y)$$

But this is Naïve Bayes!!

We are making an assumption to factorise our joint probability distribution

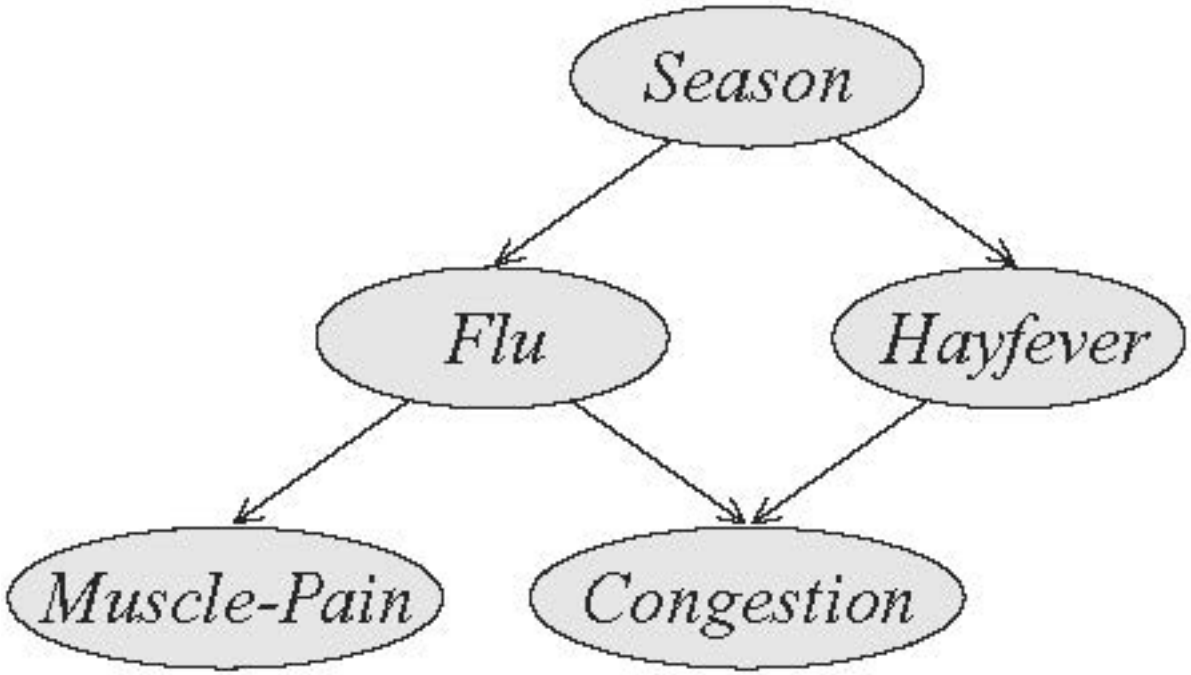
Conditional probability tables (CPTs)



Learn these tables from data

(And the structure)

Back to our patients…



If they have *Congestion*, determine likely cause from CPTs
We can get more information from *Muscle-Pain*

Testing

- Confusion matrix:

This means that 3 examples that were actually cats were incorrectly misclassified as dogs

We want these diagonal elements as high as possible, and the others as low as possible

		Actual class		
		Cat	Dog	Rabbit
Predicted class	Cat	5	2	0
	Dog	3	3	2
	Rabbit	0	1	11

$$Accuracy = \frac{diagonals}{total} = \frac{5+3+11}{5+2+3+3+2+1+11} = \frac{19}{27} = 0.704$$

This is evaluated using the test data, NOT the training data!

Verdict

- Independence assumption is very naïve:
  - Usually doesn't hold, but still useful
  - Need more sophisticated Bayesian methods (not in this course)

- But:
  - Easy to program. Simple to understand.
  - Fast to train and use.
  - Probabilistic: can deal with uncertainty