Statistical Methods in AI (CSE 471)
The Generative Way

Vineet Gandhi
Centre for Visual Information Technology (CVIT)

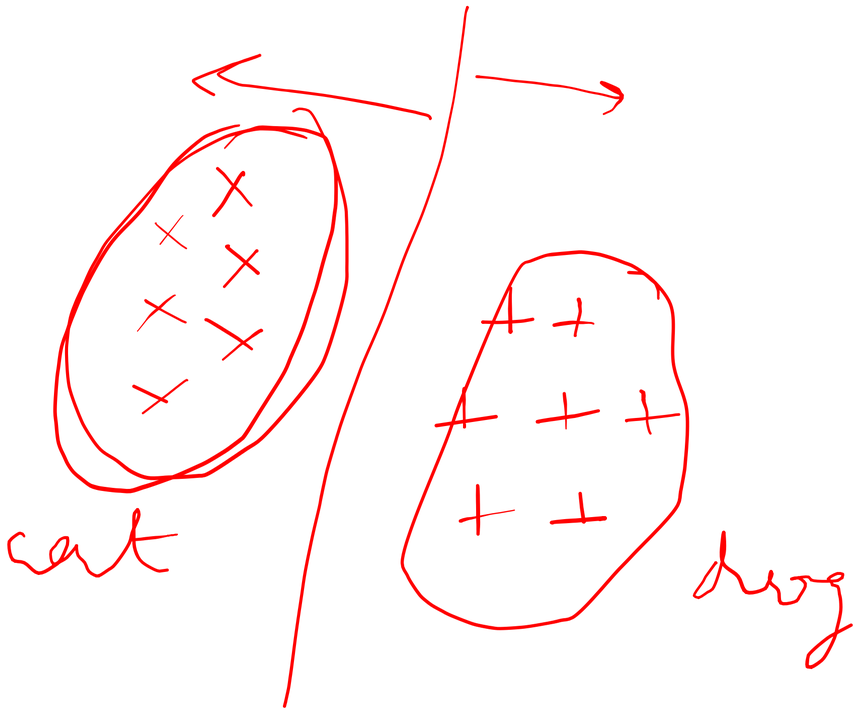
Many slides, figures, examples from: Ravi Kiran, Victor Lavrenko, Sho Nakagome, Andrew Ng, Ben Tasker, Amos Storkey

Overview

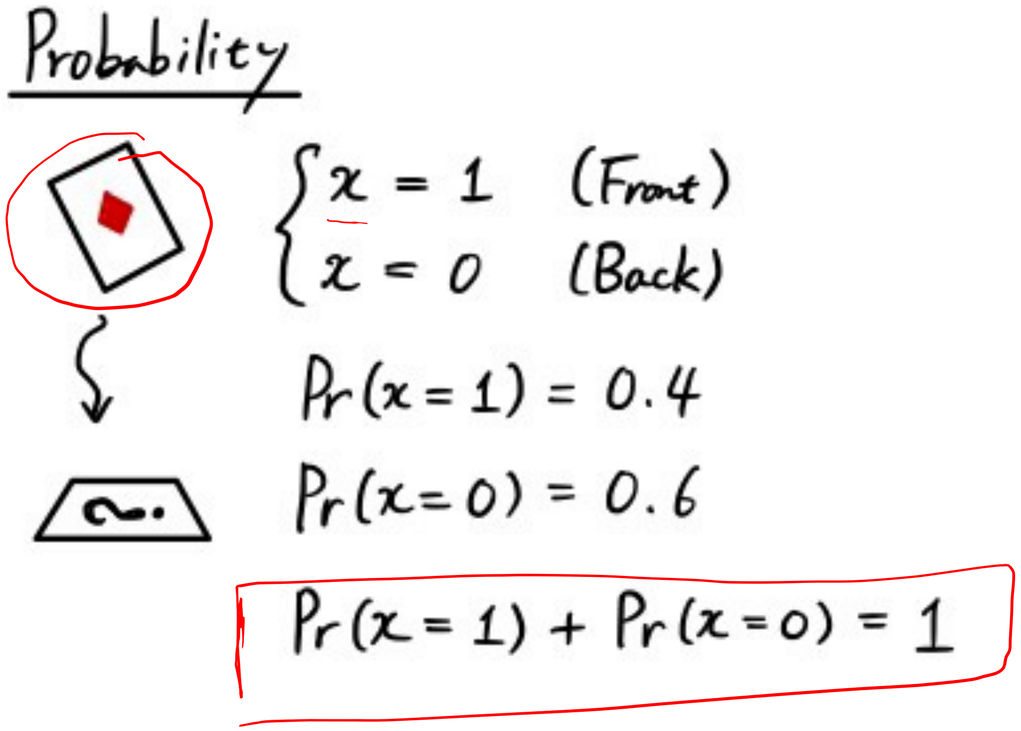
- Basics
- Gaussian Distribution - univariate + multivariate
- How to generate gaussian distribution
- Density estimation + Anomaly detection
- Bayesian Classification + Naive Bayes

Generative vs Discriminative way

- When we estimate $P(X, Y) = P(X|Y)P(Y)$, then we call it *generative learning.*
- When we only estimate $P(Y|X)$ directly, then we call it *discriminative learning.*



Probability and Random Variables



- **Random Variable (RV)** is a variable representing the outcome of our interest
- RV of x is used to represent the state of a card. It only has 2 states, either facing up (x = 1) or facing down (x = 0)
- RV could change it's state every time the event occurs. So if you dropped a card 5 times, x could be [0, 0, 1, 0, 1]
- Pr(x=0) = 3/5 = 0.6

Joint probability

$p(x, y)$

Joint Probability

$\begin{cases} x = 1 & \text{(Rains)} \\ x = 0 & \text{(Doesn't rain)} \end{cases}$

$\begin{cases} y = 1 & \text{(Have umbrella)} \\ y = 0 & \text{(Don't have umbrella)} \end{cases}$

$Pr(x = 1) = 0.6$

$Pr(x = 0) = 0.4$

$Pr(y = 1) = 0.3$

$Pr(y = 0) = 0.7$

Case 1 : Rains but you have an umbrella

$$Pr(x = 1, y = 1) = Pr(x = 1) \times Pr(y = 1)$$
$$= 0.6 \times 0.3$$
$$= 0.18$$

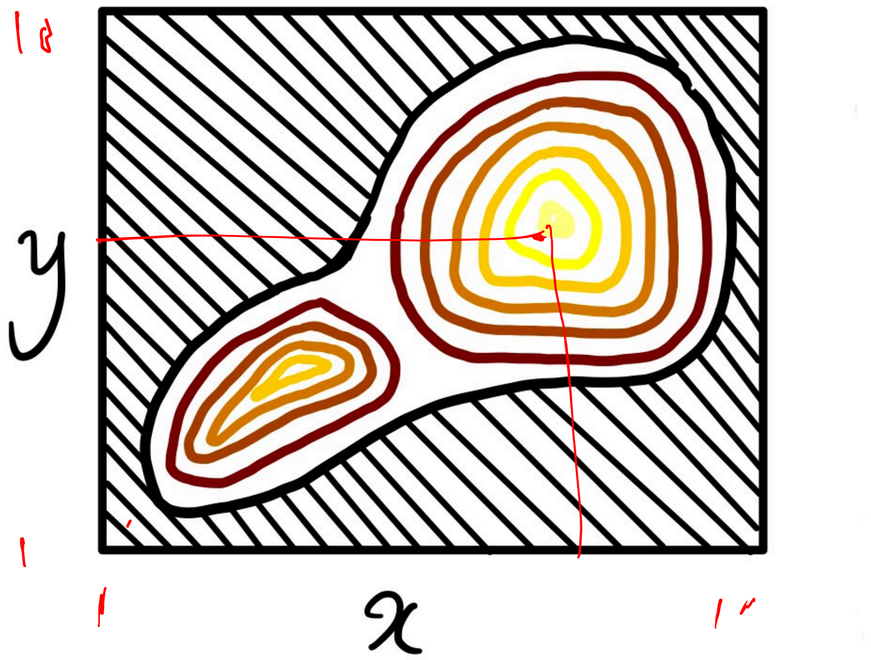
Case 2 : Rains but you DON'T have an umbrella

$$Pr(x = 1, y = 0) = Pr(x = 1) \times Pr(y = 0)$$
$$= 0.6 \times 0.7$$
$$= 0.42$$

Joint probability

$p(x=1, y=1) = \frac{\#(x=1 \,\&\, y=1)}{\#\text{obs}}$

Joint Probability in 2D



Marginalization

Marginalization

$$\begin{cases} Pr(x=0, y=0) = 0.28 \\ Pr(x=1, y=0) = 0.42 \\ Pr(x=0, y=1) = 0.12 \\ Pr(x=1, y=1) = 0.18 \end{cases}$$

How to get $Pr(x=0)$?

$$Pr(x) = \int Pr(x, y)\, dy \quad \text{(Continuous)}$$

$$Pr(x) = \sum_{y} Pr(x, y) \quad \text{(Discrete)}$$

$$Pr(x=0) = \sum_{y=0}^{1} Pr(x=0, y)$$

$$= Pr(x=0, y=0) + Pr(x=0, y=1)$$

$$= 0.28 + 0.12 = 0.4$$

Conditional probability

$$p(x|y)\,p(y) = p(y|x)\,p(x)$$

$$\Rightarrow \boxed{p(x|y) = \frac{p(y|x)\,p(x)}{p(y)}}$$

Given

$x = 1$ (Rains)

↓

What's the Probability of

$y = 1$ (Bring umbrella)

Given y,

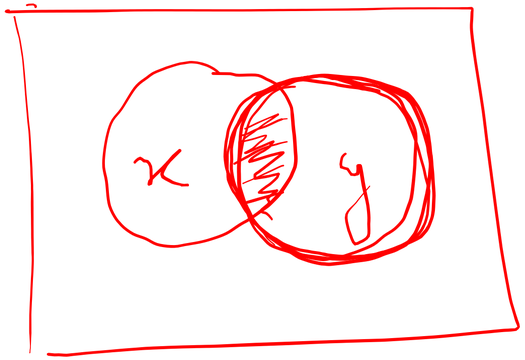
probability of x

↓

$$Pr(x|y) = \frac{Pr(x,y)}{Pr(y)}$$

Joint probability of x & y ↓

↑ Probability of y (Normalization Term)



$$p(x|y) = \frac{p(x,y)}{p(y)}$$

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

Marginalization

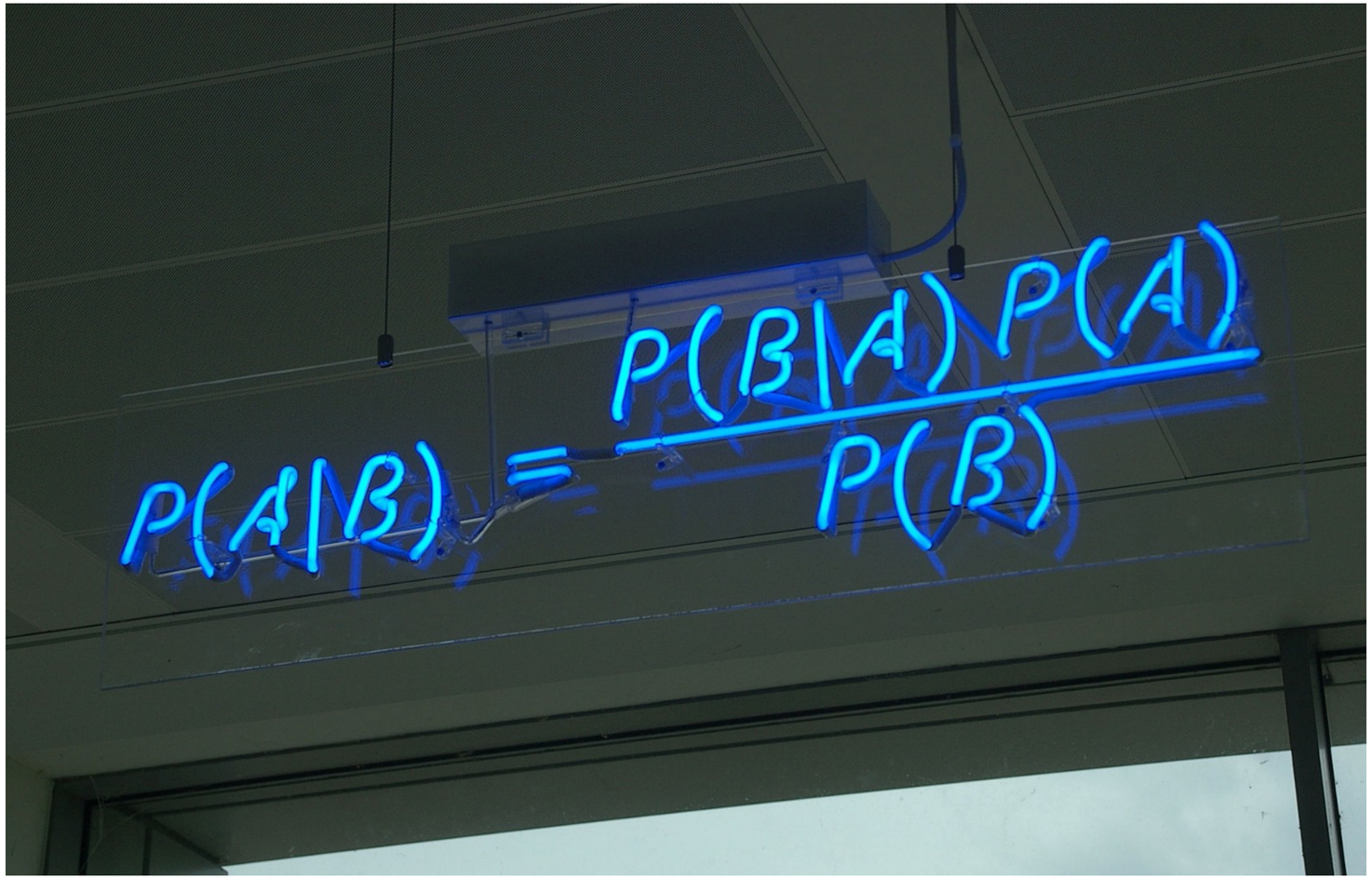
$$p(x=0) = p(x=0 \mid y=0) + p(x=0 \mid y=1)$$

$$p(x) = p(x,a) + p(x,b)$$

$$p(x) = p(x \mid a)\,p(a) + p(x \mid b)\,p(b)$$

$$p(x=0) = p(x=0 \mid y=0) \cdot p(y=0) + p(x=0 \mid y=1)\,p(y=1)$$

Bayes' Rule



- A disease occurs in 0.5% of population
- A diagnostic test gives a positive result
  - in 99% of people that have the disease
  - in 5% of people that do not have the disease (false positive)

A random person from the street is found to be positive on this test. What is the probability that they have the disease?

A: 0-30%

B: 30-60%

C: 60-90%

- A disease occurs in 0.5% of population
- A diagnostic test gives a positive result
  - in 99% of people that have the disease
  - in 5% of people that do not have the disease (false positive)

A = disease

B = positive test result

P(A) = 0.005 probability of having disease

P(~A) = 1 – 0.005 = 0.995 probability of not having disease

P(B) = 0.005 * 0.99 (people with disease) + 0.995 * 0.05 (people without disease) = 0.0547 (slightly more than 5% of *all* tests are positive)

conditional probabilities

P(B|A) = 0.99 probability of pos result **given** you have disease

P(~B|A) = 1 – 0.99 = 0.01 probability of neg result **given** you have disease

P(B|~A) = 0.05 probability of pos result **given** you do not have disease

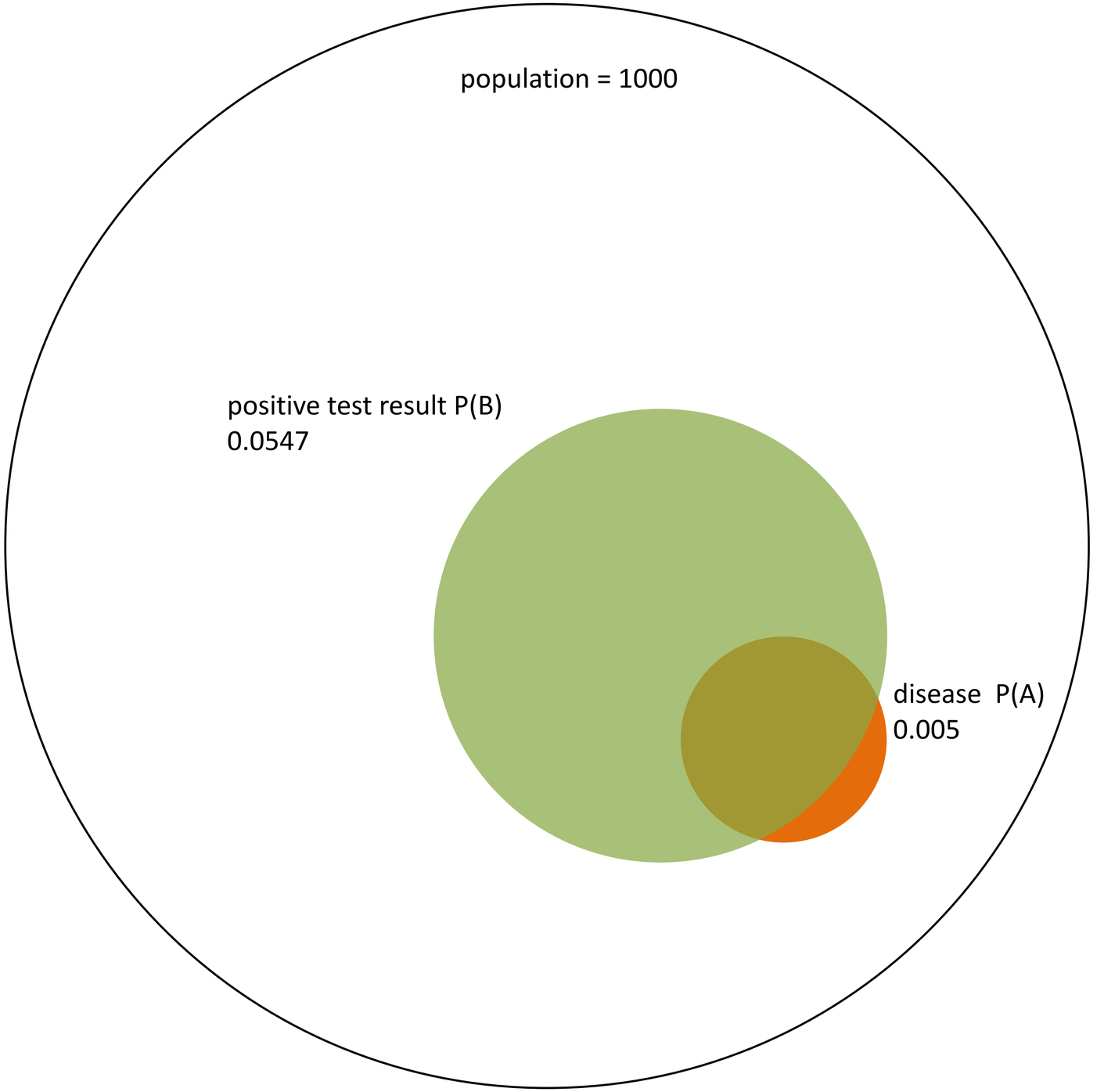
P(~B|~A) = 1 – 0.05 = 0.95 probability of neg result **given** you do not have disease

P(A|B) is probability of disease *given* the test is positive (which is what we're interested in)

Very different from P(B|A): probability of positive test results given you have the disease.

A = disease

B = positive test result



A = disease

B = positive test result

P(A,B) is the *joint probability*, or the probability that both events occur. P(A,B) is the same as P(B,A).

But we already *know* that the test was positive, so we have to take that into account.

Of all the people already in the green circle, how many fall into the P(A,B) part? That's the probability we want to know!

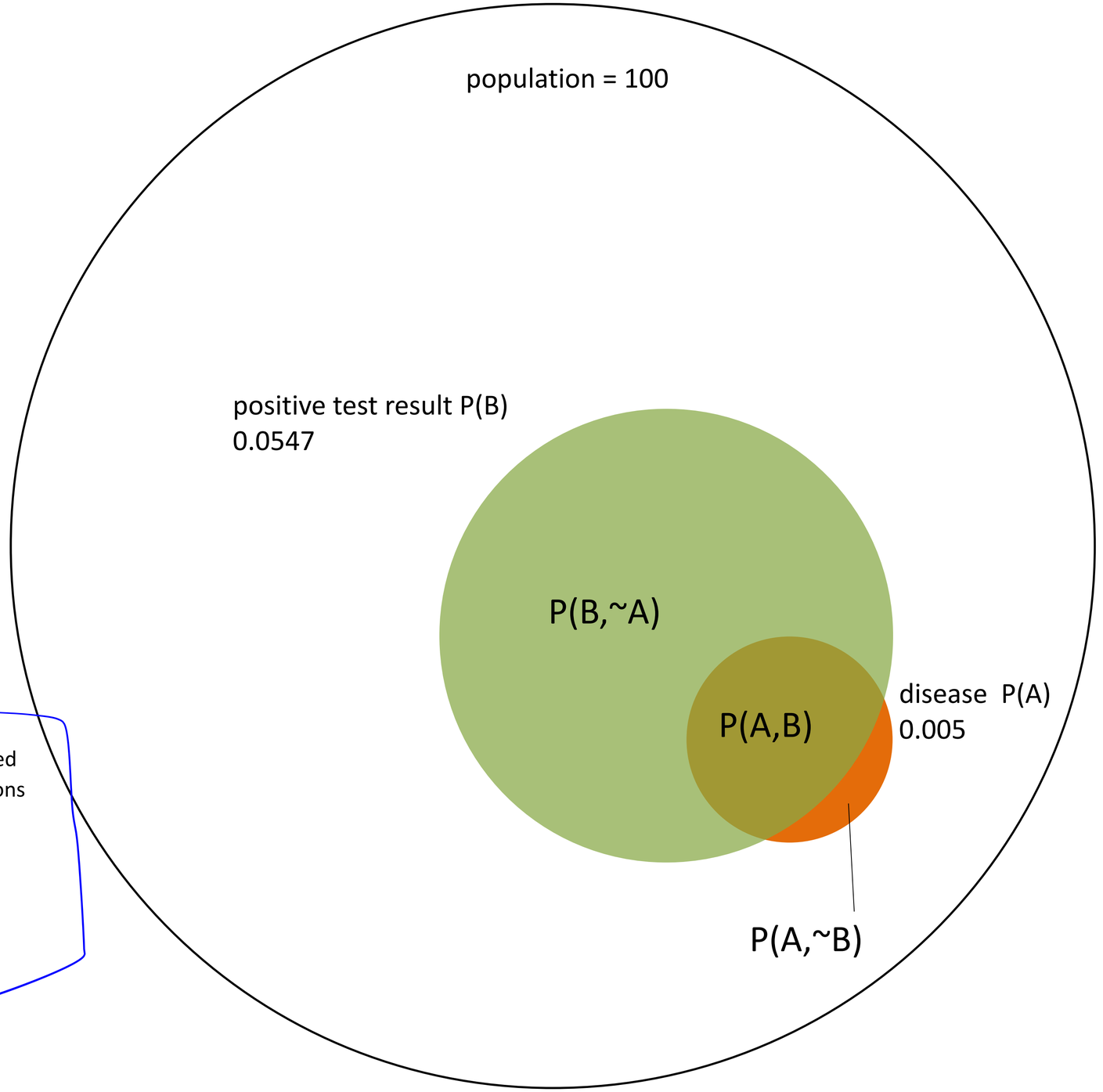
That is:
P(A|B) = P(A,B) / P(B)

You can write down same thing for the inverse:
P(B|A) = P(A,B) / P(A)

The *joint probability* can be expressed in two ways by rewriting the equations
P(A,B) = P(A|B) * P(B)
P(A,B) = P(B|A) * P(A)

Equating the two gives
P(A|B) * P(B) = P(B|A) * P(A)

P(A|B) = P(B|A) * P(A) / P(B)



A = disease

B = positive test result

P(A) = 0.005 probability of having disease

P(B|A) = 0.99 probability of pos result **given** you have disease

P(B) = 0.005 * 0.99 (people with disease) + 0.995 * 0.05 (people without disease) = 0.0547

Bayes' Theorem

P(A|B) = P(B|A) * P(A) / P(B)

P(A|B) = 0.99 * 0.005 / 0.0547

= 0.09

So a positive test result increases your probability of having the disease to 'only' 9%, simply because the disease is very rare (relative to the false positive rate).

P(A) is called the **prior**: before we have any information, we estimate the chance of having the disease 0.5%

P(B|A) is called the **likelihood**: probability of the data (pos test result) given an underlying cause (disease)

P(B) is the **marginal probability of the data:** the probability of observing this particular outcome, taken over all possible values of A (disease and no disease)

P(A|B) is the **posterior probability**: it is a combination of what you thought before obtaining the data, and the new information the data provided (combination of **prior** and **likelihood**)

Let's do another one…

It rains on 20% of days.
When it rains, it was forecasted 80% of the time
When it doesn't rain, it was erroneously forecasted 10% of the time.

The weatherman forecasts rain. What's the probability of it actually raining?

A = forecast rain
B = it rains

What information is given in the story?

P(B) = 0.2 (**prior**)
P(A|B) = 0.8 (**likelihood**)
P(A|~B) = 0.1

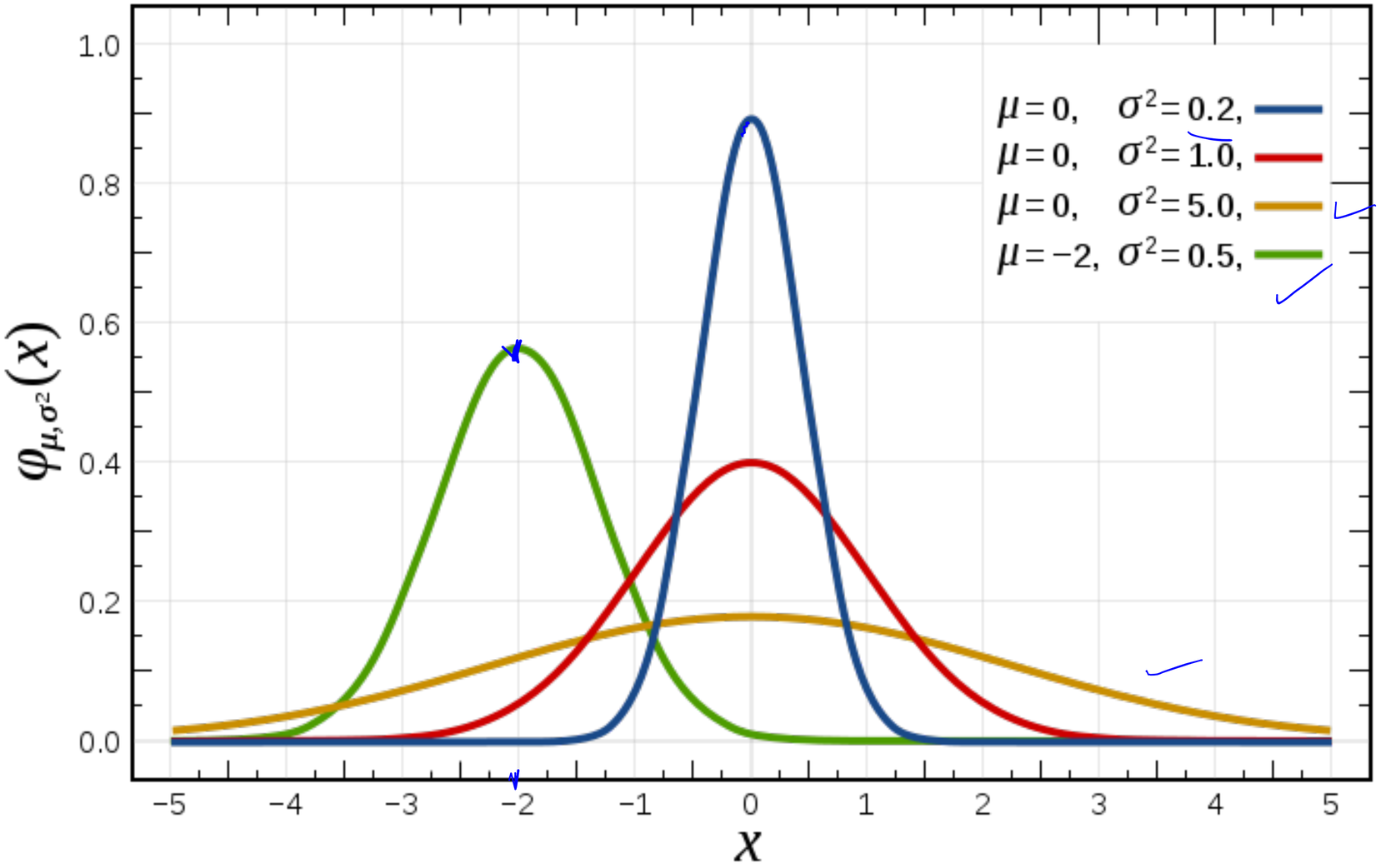
P(B|A) = P(A|B) * P(B) / P(A)

What is P(A), probability of rain forecast? Calculate over all possible values of B (**marginal probability**)
P(A|B) * P(B) + P(A|~B) * P(~B) = 0.8 * 0.2 + 0.1 * 0.8 = 0.24

P(B|A) = 0.8 * 0.2 / 0.24
= 0.67

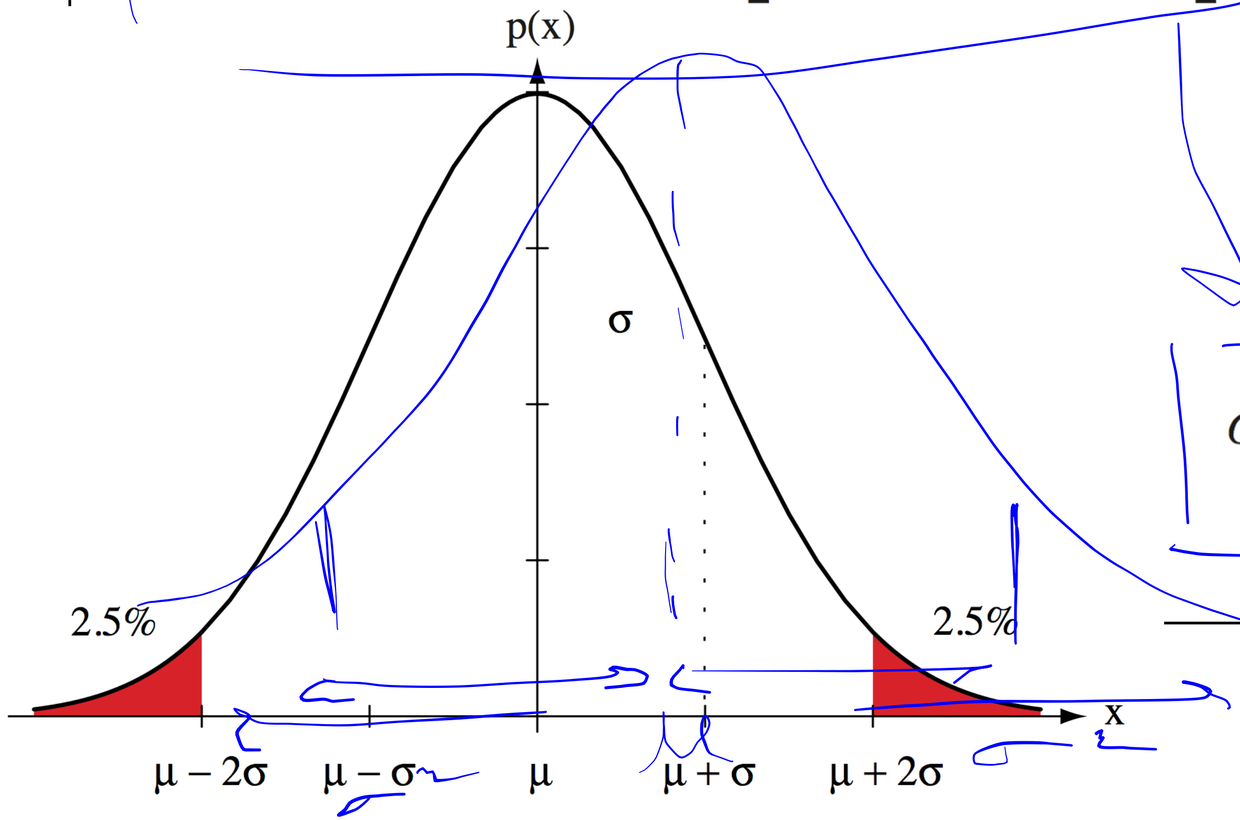
So before you knew anything you thought P(rain) was 0.2. Now that you heard the weather forecast, you adjust your expectation upwards P(rain|forecast) = 0.67

Univariate Gaussian Distribution



Parameter estimation

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$



$$\mu \equiv \mathcal{E}[x] = \int_{-\infty}^{\infty} x p(x)\, dx$$

$$\sigma^2 \equiv \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx$$

Anomaly Detection

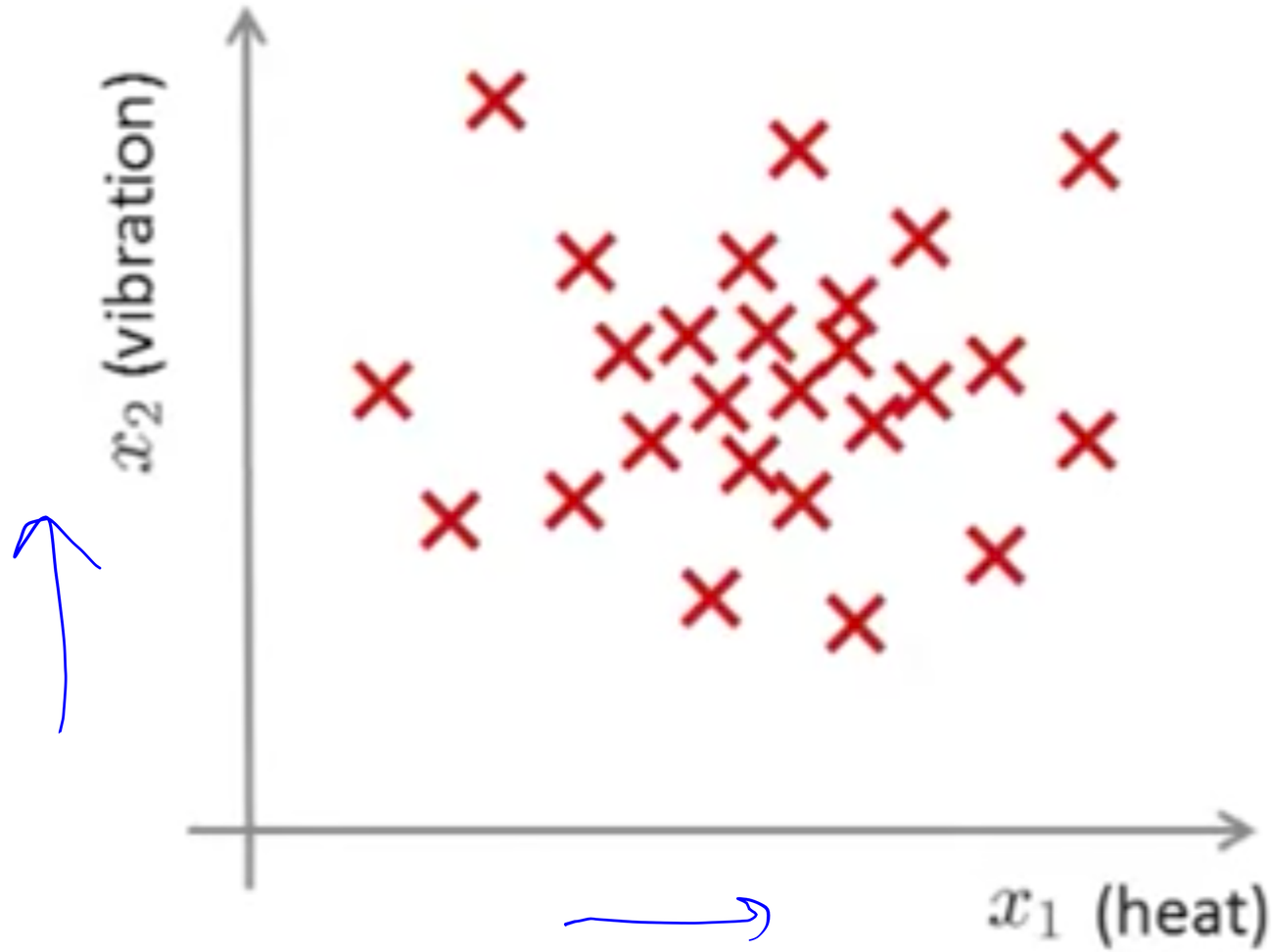


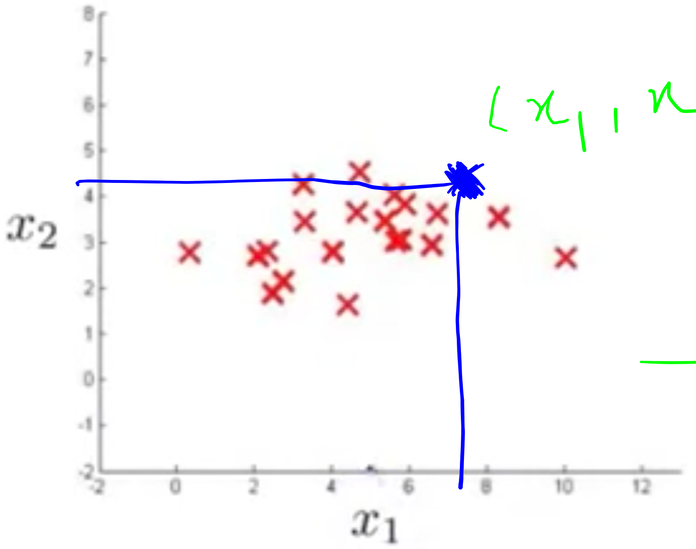
figure credit: Andrew Ng

Density Estimation

$p(x) = p(x_1, x_2, x_3, \ldots, x_d) = \ldots$

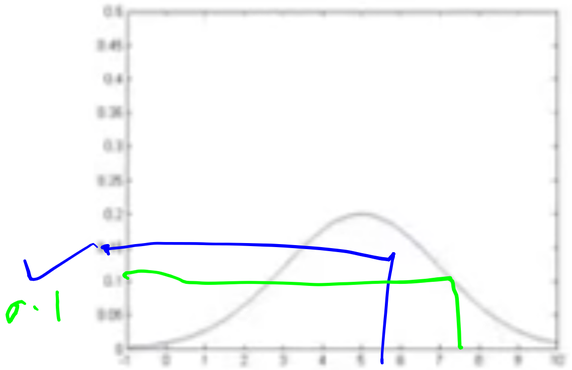
$= p(x_1 \mid \mu_1, \sigma_1^2) * p(x_2 \mid \mu_2, \sigma_2^2) \cdot p(x_3 \mid \mu_3, \sigma_3)$

Example: Anomaly Detection

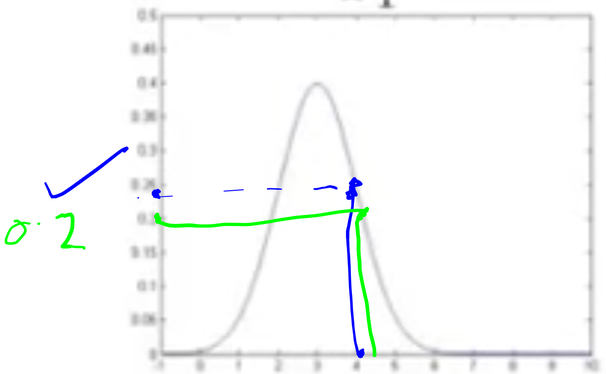


$$\mu_1 = 5, \sigma_1 = 2$$

$$\mu_2 = 3, \sigma_2 = 1$$



$$p(x_1; \mu_1, \sigma_1^2)$$



$$p(x_2; \mu_2, \sigma_2^2)$$

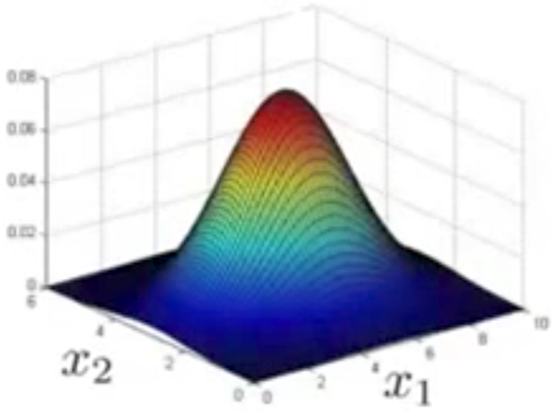
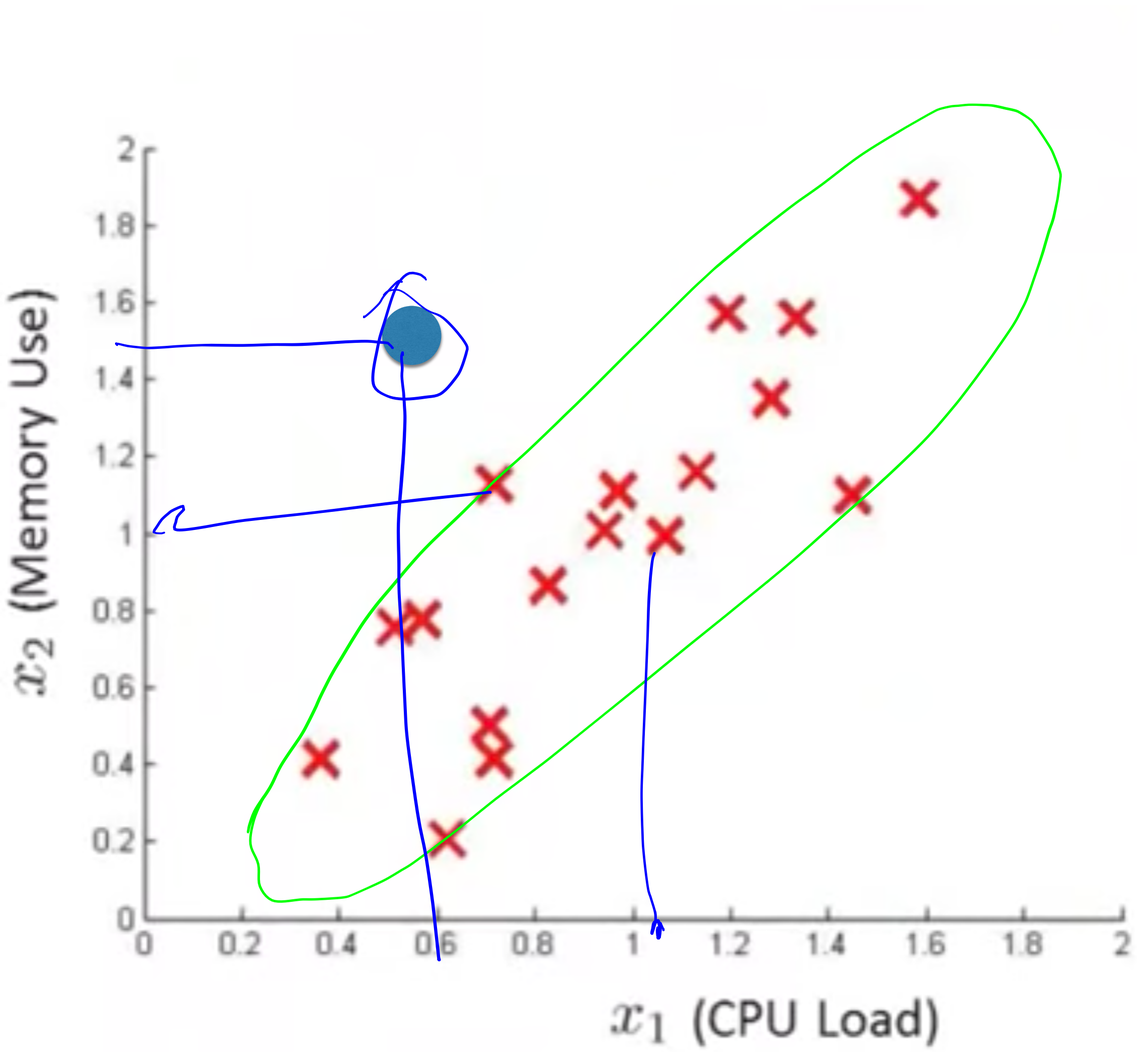
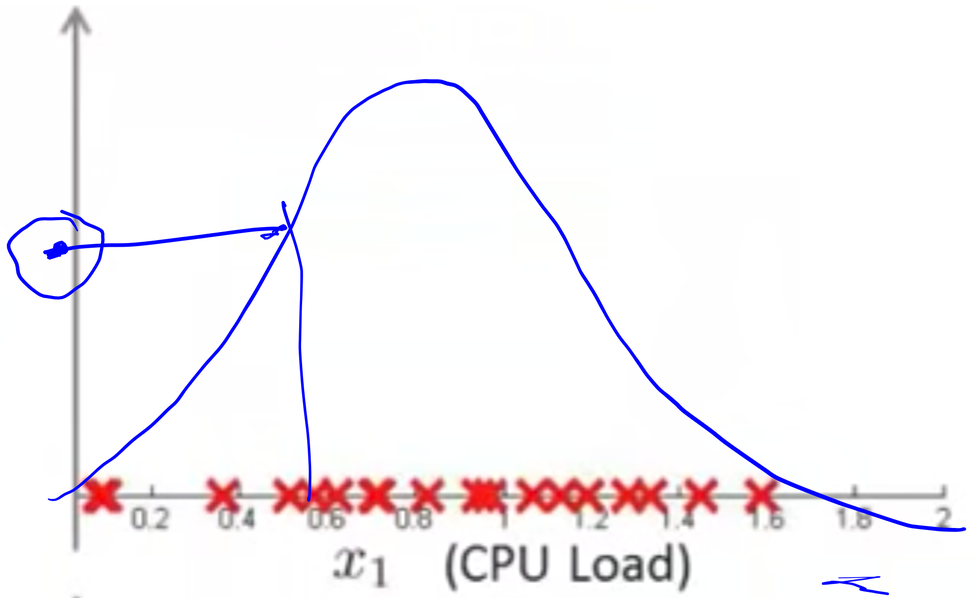


figure credit: Andrew Ng

Multivariate Gaussian Distribution





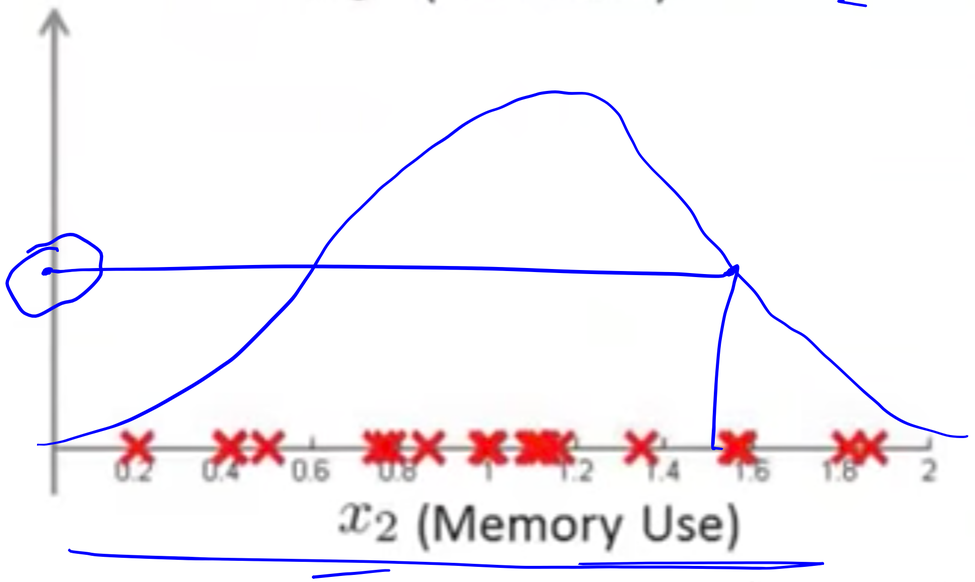


figure credit: Andrew Ng

Multivariate Gaussian Distribution

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

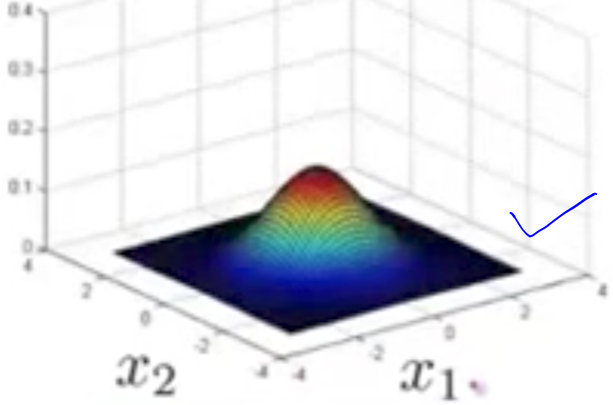
$$\boldsymbol{\mu} \equiv \mathcal{E}[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x})\, d\mathbf{x}$$

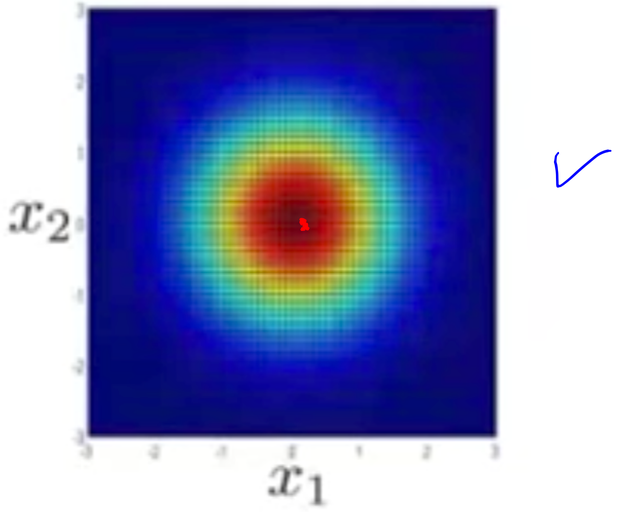
$$\mathbf{\Sigma} \equiv \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t p(\mathbf{x})\, d\mathbf{x},$$

$$\mu_i = \mathcal{E}[x_i] \qquad \sigma_{ij} = \mathcal{E}[(x_i-\mu_i)(x_j-\mu_j)].$$

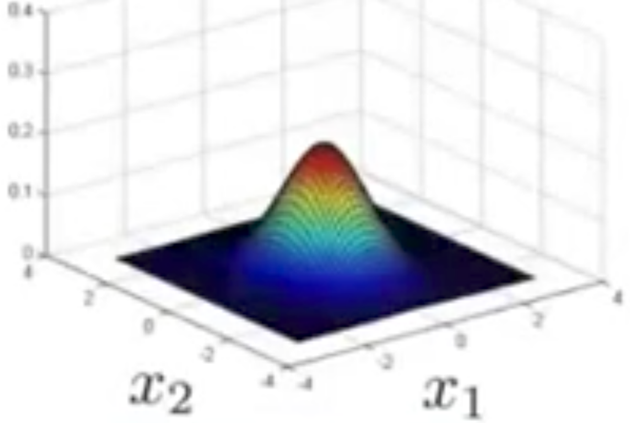
Multivariate Gaussian Distribution

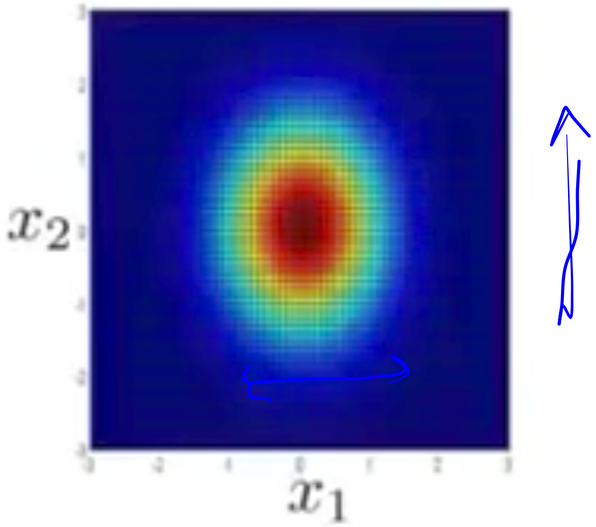
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



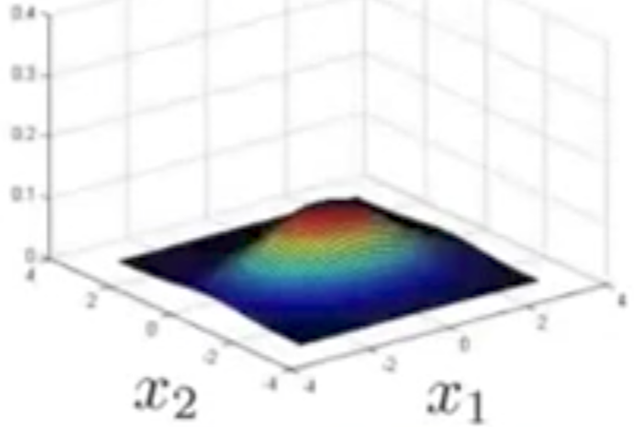


$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$$





$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$



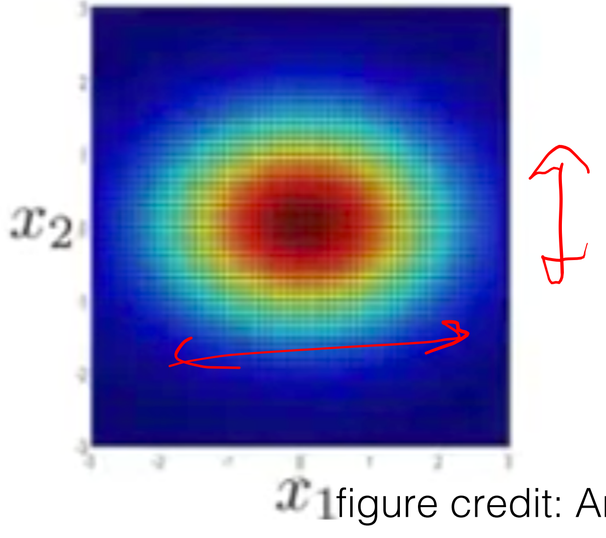
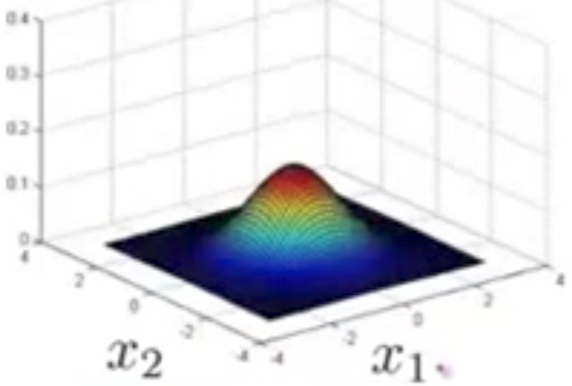
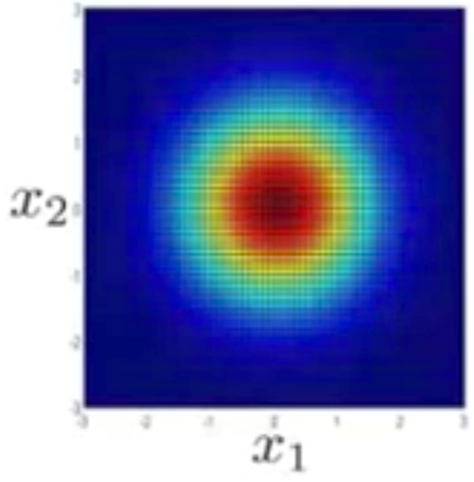


figure credit: Andrew Ng

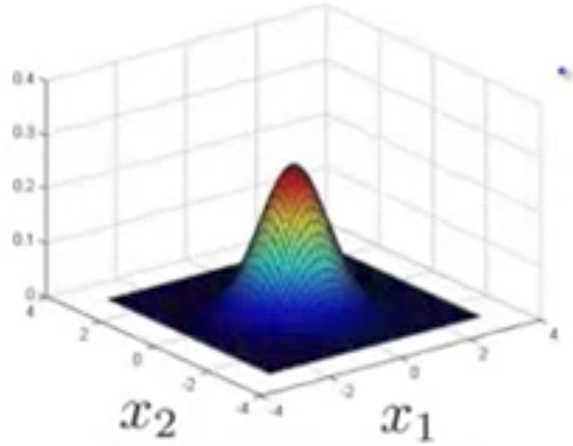
Multivariate Gaussian Distribution

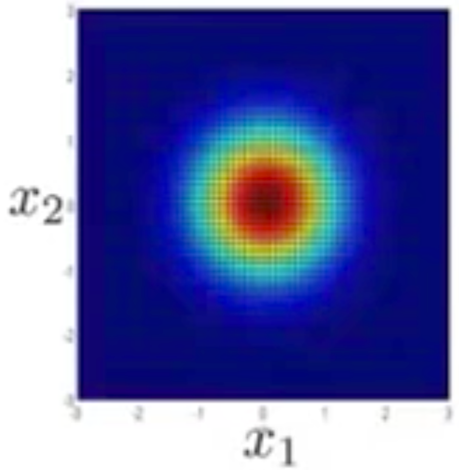
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



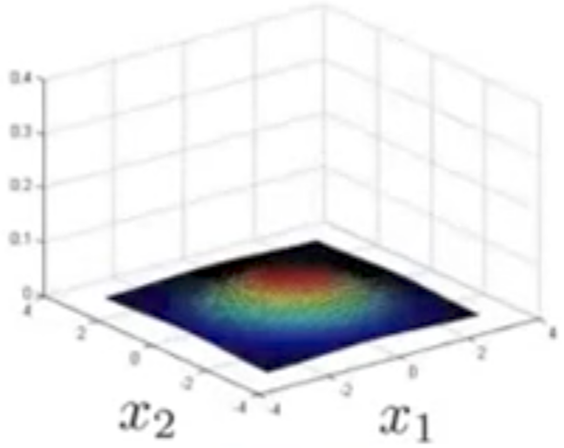


$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$$





$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$



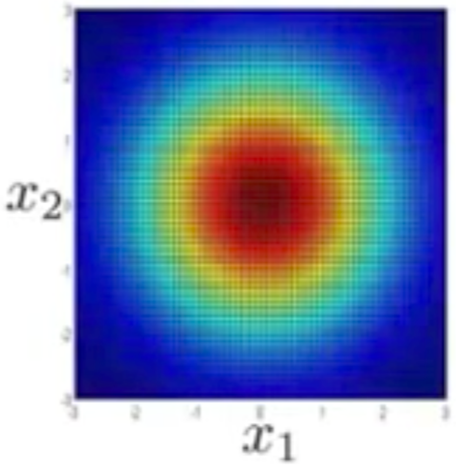
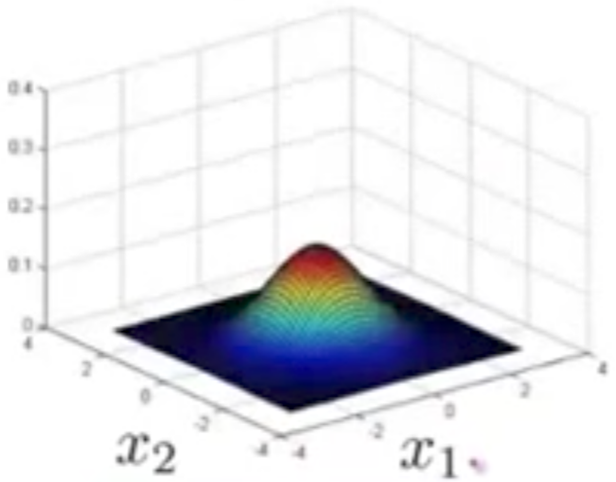
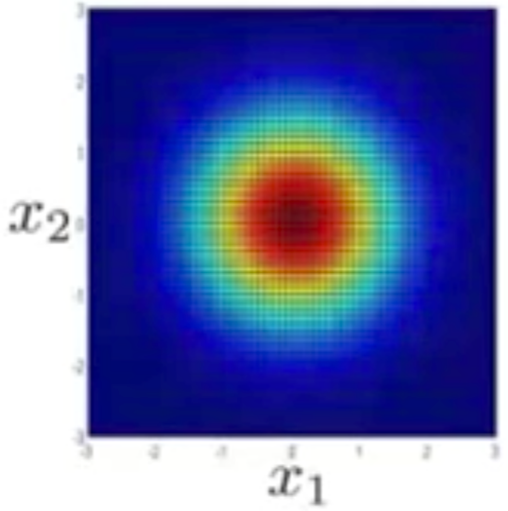


figure credit: Andrew Ng

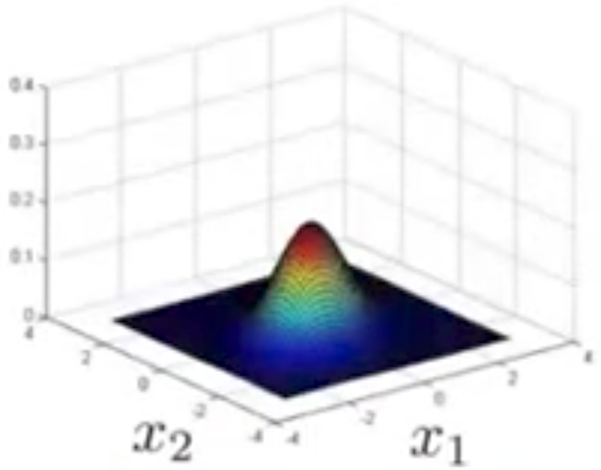
Multivariate Gaussian Distribution

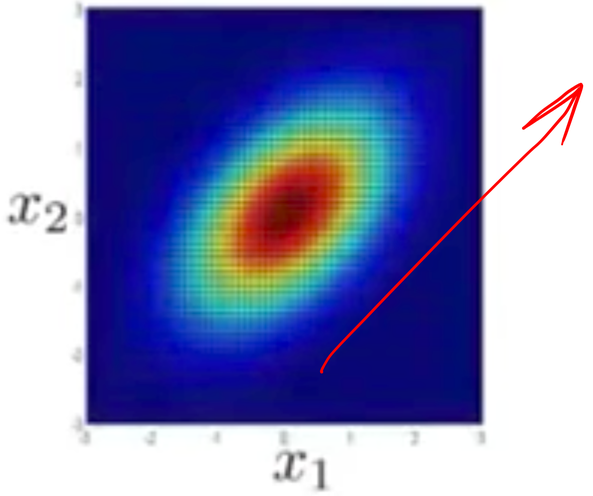
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



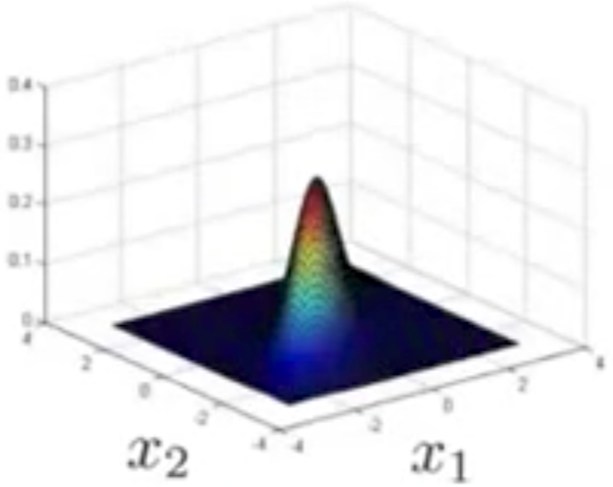


$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$





$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



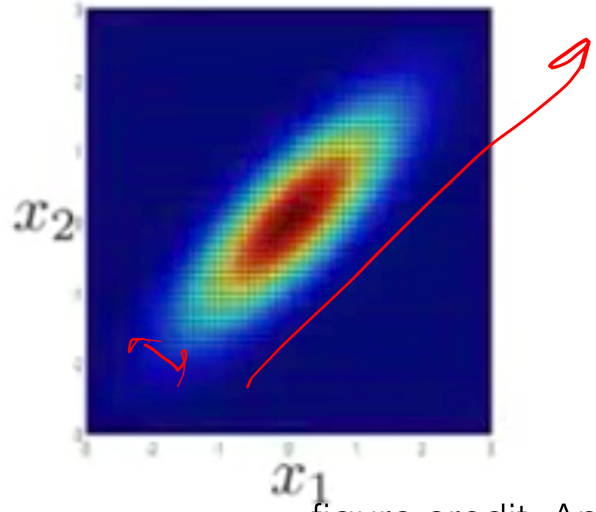
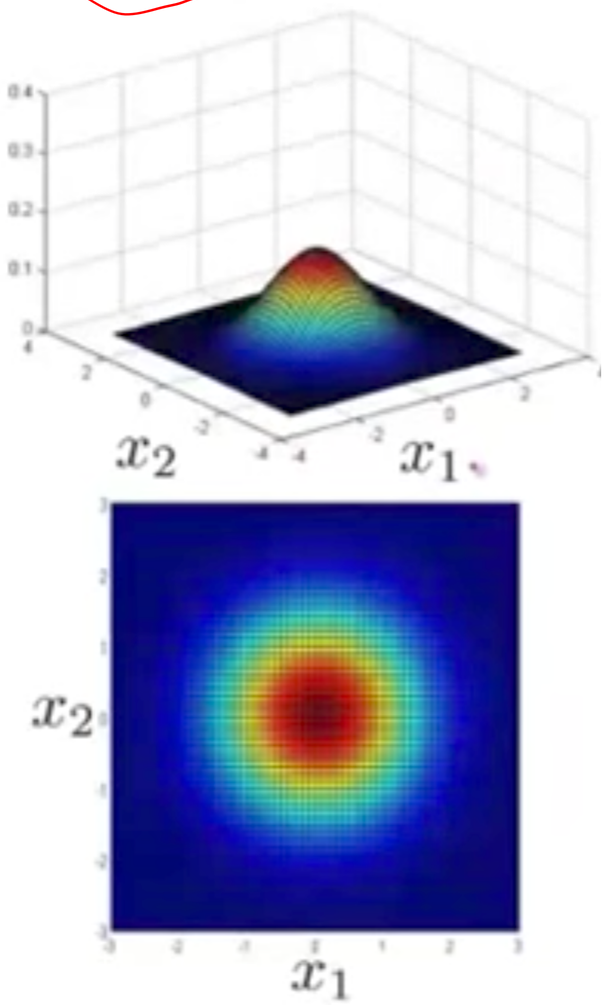


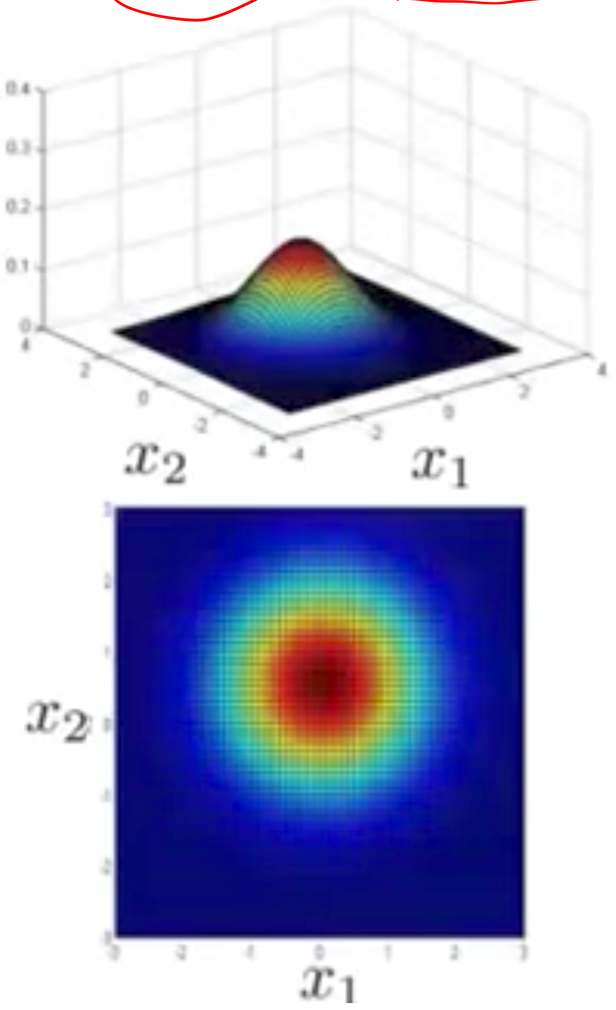
figure credit: Andrew Ng

Multivariate Gaussian Distribution

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

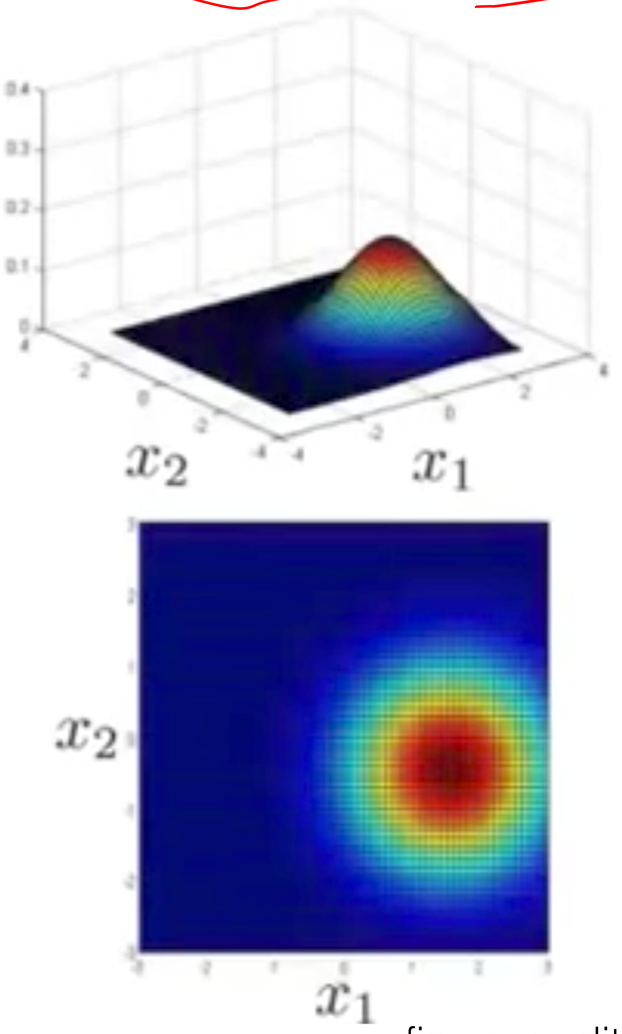


figure credit: Andrew Ng

Anomaly detection: using multivariate Gaussian distribution

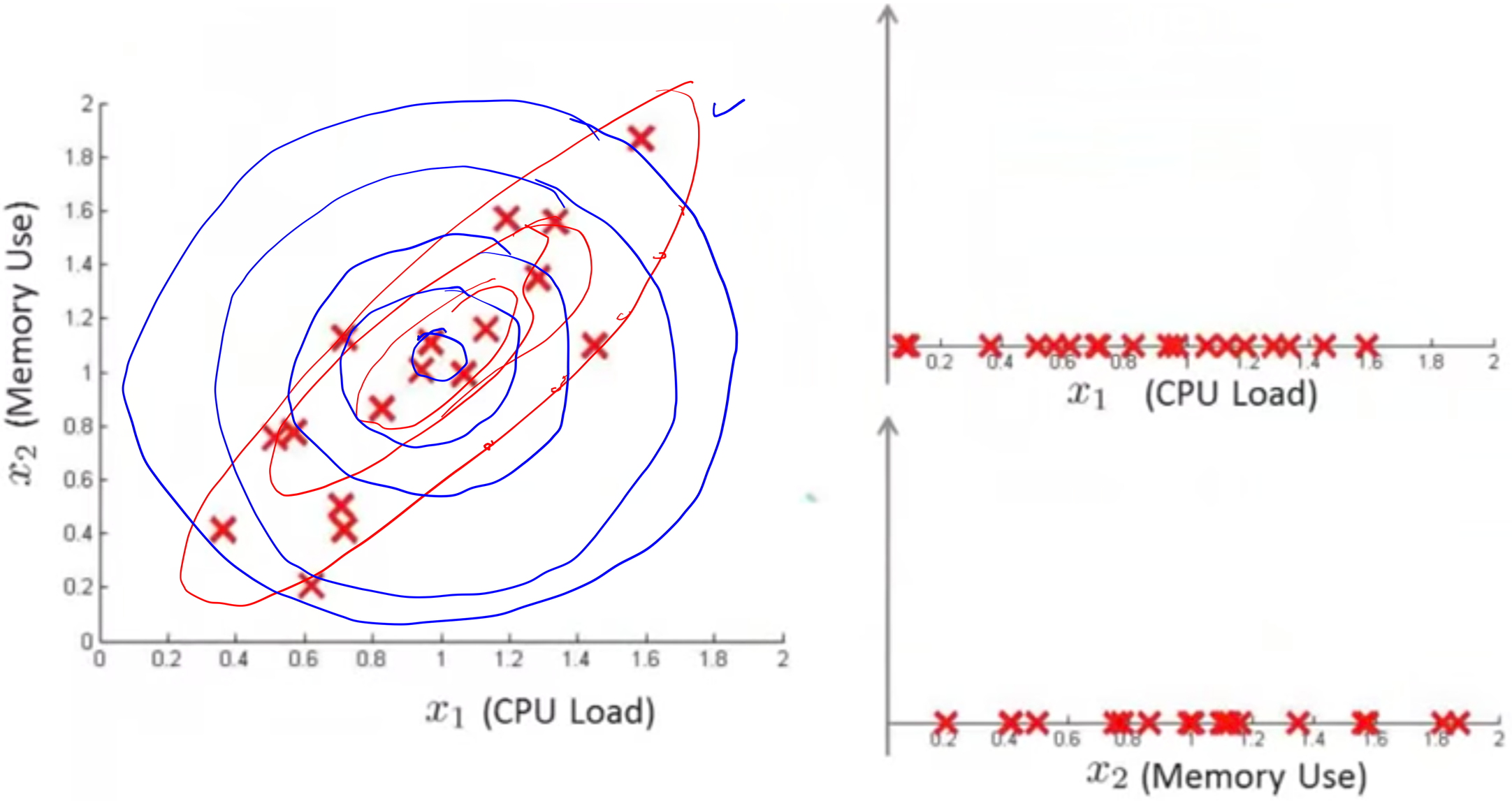


figure credit: Andrew Ng

Anomaly detection	vs.	Supervised learning
Very small number of positive examples ($y = 1$). (0-20 is common).		Large number of positive and negative examples.
Large number of negative ($y = 0$) examples.		
Many different "types" of anomalies. Hard for any algorithm to learn from positive examples what the anomalies look like; future anomalies may look nothing like any of the anomalous examples we've seen so far.		Enough positive examples for algorithm to get a sense of what positive examples are like, future positive examples likely to be similar to ones in training set.

figure credit: Andrew Ng

Bayesian Classification

- Goal: learning functions f(x) —> y
  - y —> one of the k classes
  - x= $x_1, x_2, \ldots x_d$ - values of attributes (numerical or categorical)

- Probabilistic classification:
  - most probable class given observations: $\hat{y} = \underset{y}{\text{argmax}}\, P(y \mid x)$

- Bayesian probability of a class:

Likelihood **Prior**

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

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

Normalization —> ranking

Independence assumption: Naive Bayes

- We need to compute P(x|y) for Bayesian classification
- x is multidimensional, which makes counting like easy probability estimation difficult

Idea: Make independence assumption

$$P(x_1, \ldots x_d|y) = \prod_{i=1}^{d} P(x_i|x_1 \ldots x_{i-1}, y) = \prod_{i=1}^{d} P(x_i|y)$$

assuming that the class value explains all the dependence between attributes

Continuous Example

- Distinguish children from adults based on height and weight
  - classes: {a,c} , attributes: height (cm), weight (kg)
  - training examples: $\{h_i, w_i, y_i\}$, 4 adults and 12 children
- Prior probabilities: P(a) = 4/16 and P(c) = 12/16
- Model: gaussians using mean and variances

Continuous Example

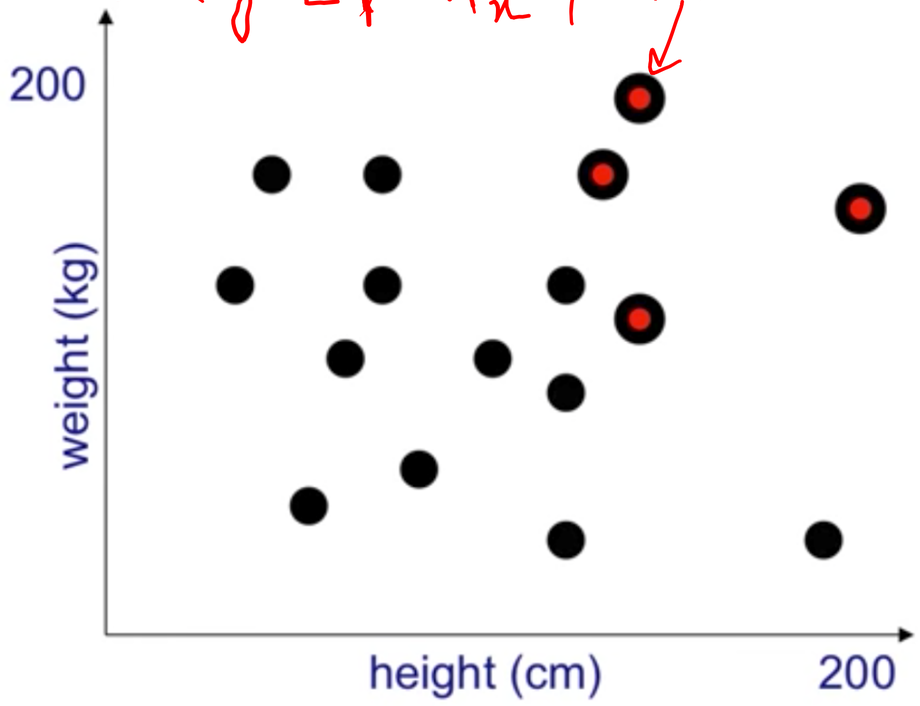
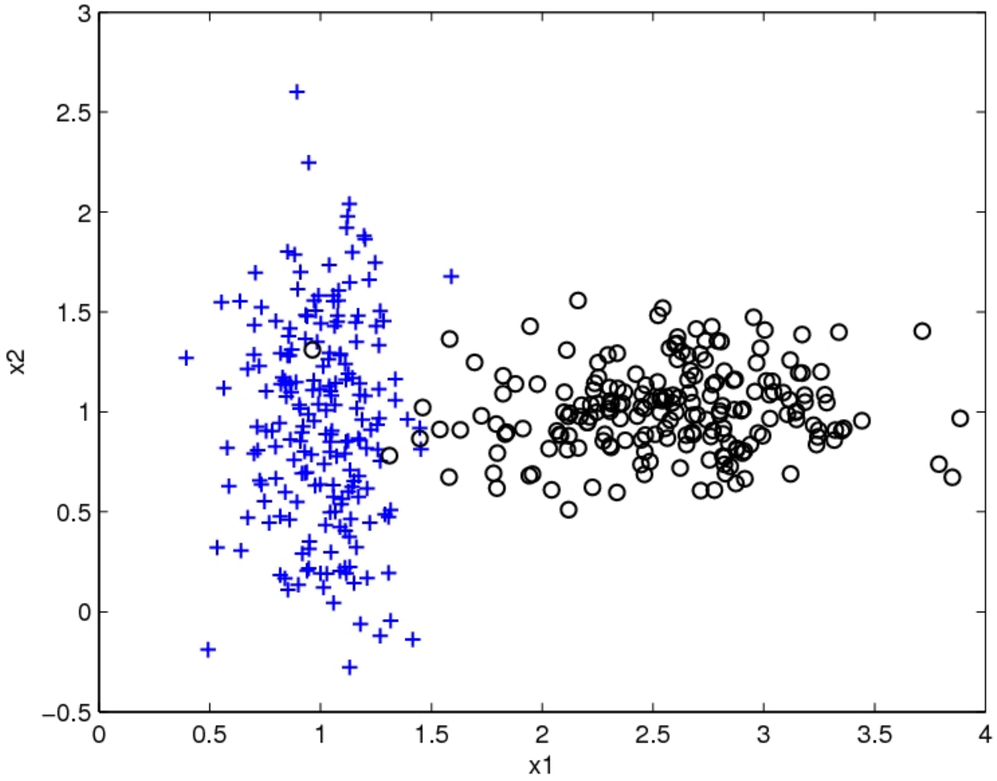


figure credit: Victor Lavrenko

Decision boundary



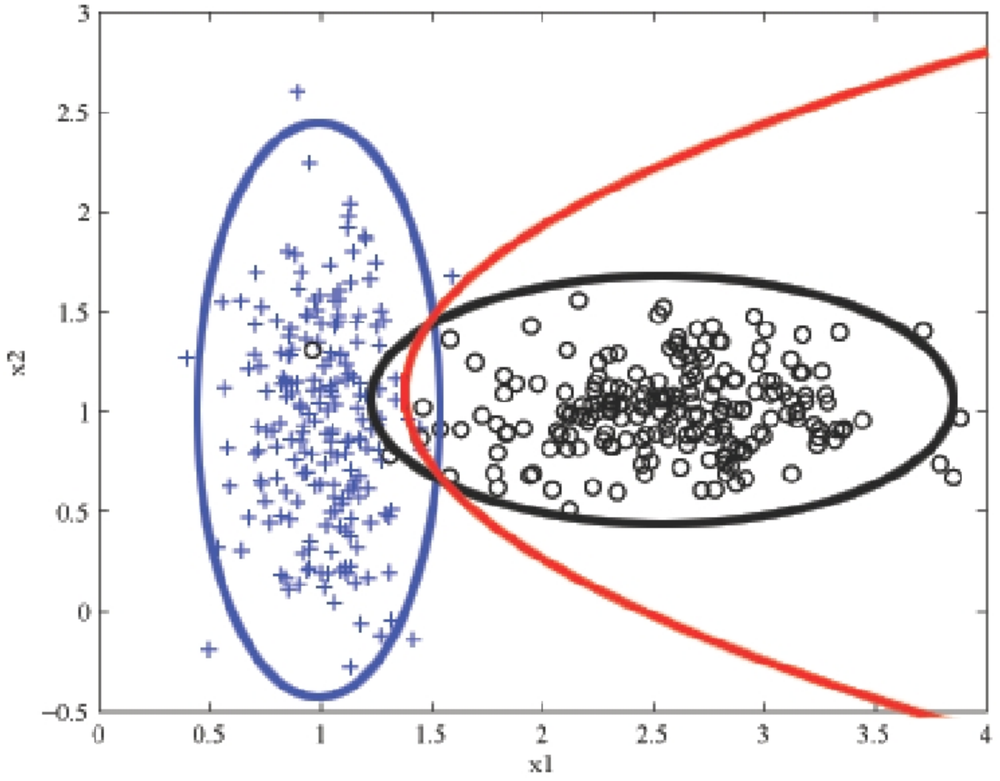


figure credit: Ben Tasker

Problem with Naive Bayes

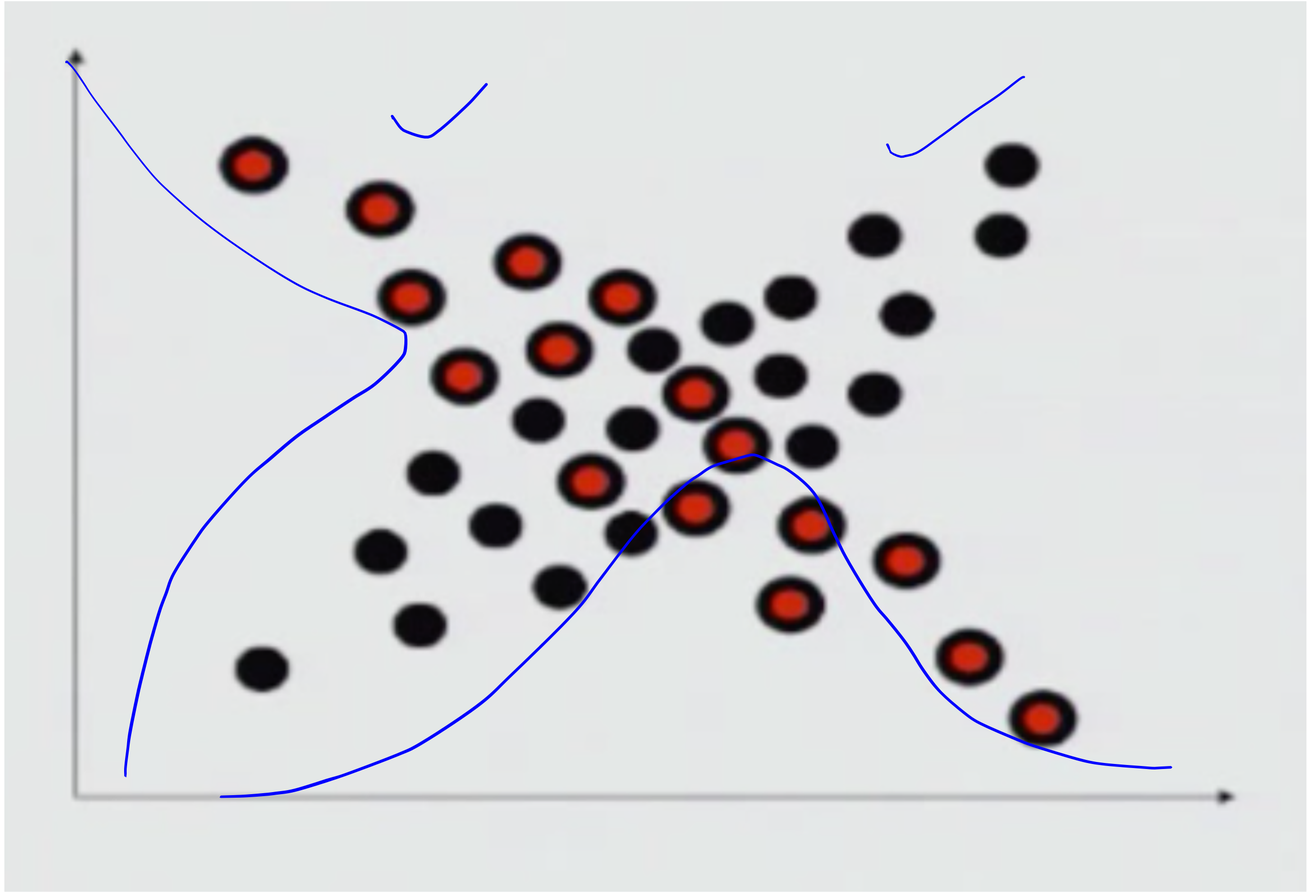


figure credit: Victor Lavrenko

Discrete example

$p(\text{spam} \mid \text{review us}) = p(\text{review us} \mid \text{spam})\, p(\text{spam})$

$p(r \mid \text{ham}) = 2/2$

$p(\text{spam}) = 4/6 \quad p(\text{ham}) = 2/6$

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

spam	ham	
2/4	1/2	password x_1
1/4	2/2	review x_2
3/4	1/2	send
3/4	1/2	us
3/4	1/2	your
1/4	0/2	account x_6

× new email: “review us now”

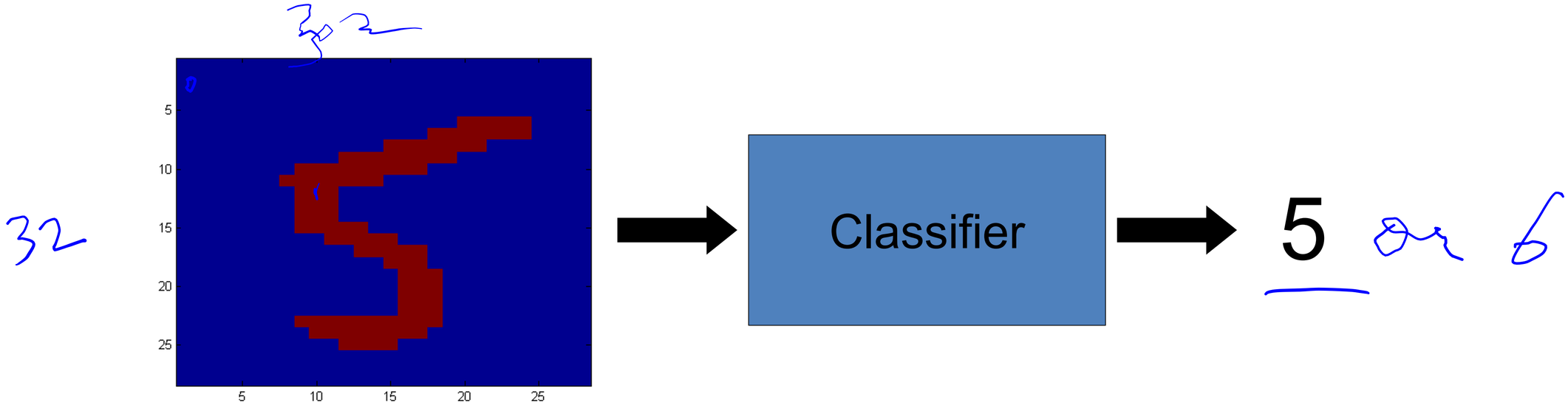
$2^{50,000}$

$p(\text{review us} \mid \text{spam}) = p(0,1,0,1,0,0 \mid \text{spam})$

$= \left(1-\frac{2}{4}\right)\left(\frac{1}{4}\right)\left(1-\frac{3}{4}\right)\left(\frac{3}{4}\right)\left(1-\frac{3}{4}\right)\left(1-\frac{1}{4}\right)$

Another Application

- **Digit Recognition**



- $X_1,\ldots,X_n \in \{0,1\}$ (Black vs. White pixels)
- $Y \in \{5,6\}$ (predict whether a digit is a 5 or a 6)

The Bayes Classifier

- A good strategy is to predict:

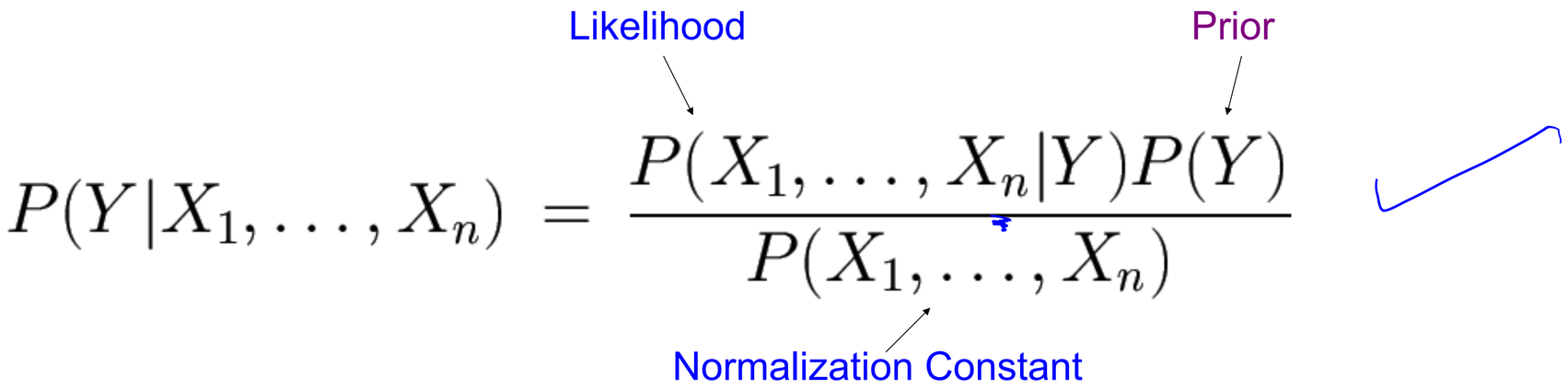
$$\arg\max_Y P(Y | X_1, \ldots, X_n)$$

  - (for example: what is the probability that the image represents a 5 given its pixels?)

- How do we compute that?

The Bayes Classifier

- Use Bayes Rule!



- Why did this help? Well, we think that we might be able to specify how features are "generated" by the class label

The Bayes Classifier

- Let's expand this for our digit recognition task:

$$P(Y=5|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y=5)P(Y=5)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}$$

$$P(Y=6|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y=6)P(Y=6)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}$$

- To classify, we'll simply compute these two probabilities and predict based on which one is greater

$2^{32\times32}$

Model Parameters

- For the Bayes classifier, we need to “learn” two functions, the likelihood and the prior

- How many parameters are required to specify the prior for our digit recognition example?

Model Parameters

- How many parameters are required to specify the likelihood?
  - (Supposing that each image is 30x30 pixels)

?

Model Parameters

- The problem with explicitly modeling $P(X_1,\ldots,X_n|Y)$ is that there are usually way too many parameters:
  - We'll run out of space
  - We'll run out of time
  - And we'll need tons of training data (which is usually not available)

The Naïve Bayes Model

- The *Naïve Bayes Assumption*: Assume that all features are independent **given the class label Y**
- Equationally speaking:

$$P(X_1, \ldots, X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

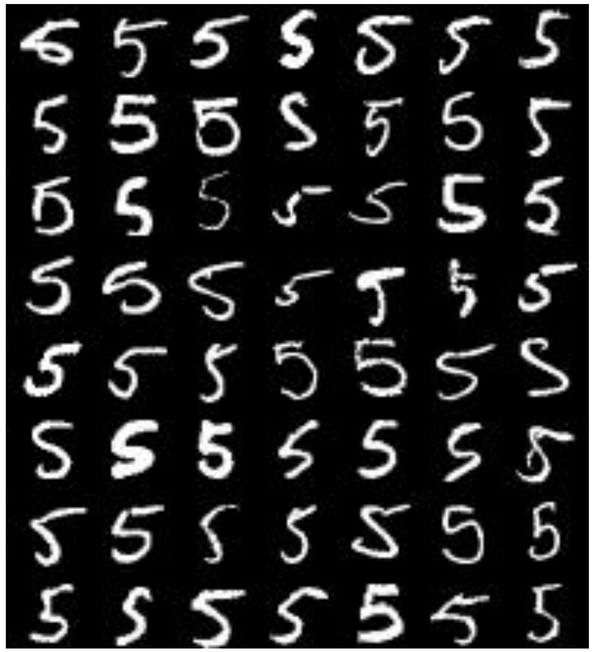
- (We will discuss the validity of this assumption later)

Why is this useful?

- # of parameters for modeling $P(X_1,\ldots,X_n|Y)$:
  - $2(2^n-1)$
- # of parameters for modeling $P(X_1|Y),\ldots,P(X_n|Y)$
  - $2n$

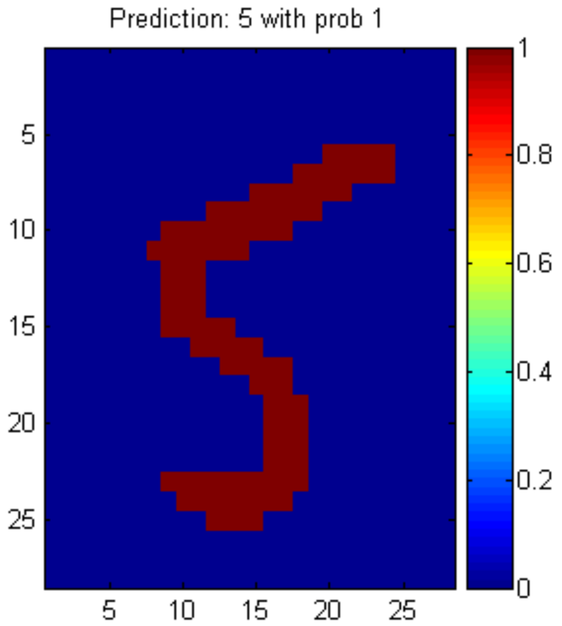
Naïve Bayes Training

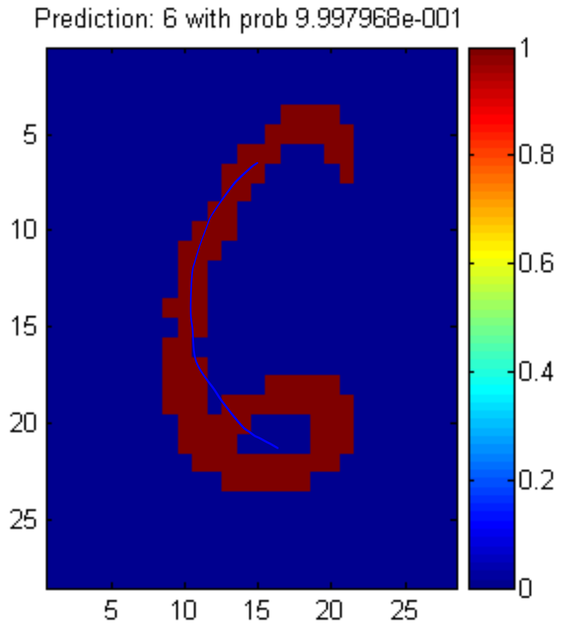
- Now that we've decided to use a Naïve Bayes classifier, we need to train it with some data:

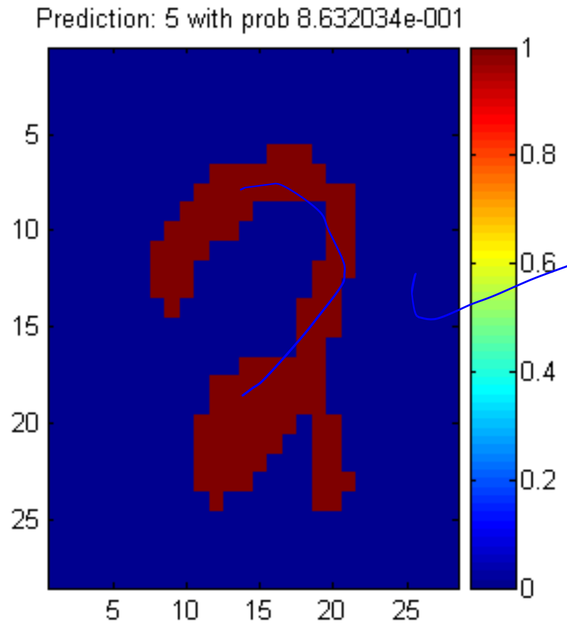


MNIST Training Data

Naïve Bayes Classification







Naïve Bayes Training

- Training in Naïve Bayes is **easy**:
  - Estimate P(Y=v) as the fraction of records with Y=v

$$P(Y = v) = \frac{Count(Y = v)}{\#\ records}$$

  - Estimate $P(X_i=u|Y=v)$ as the fraction of records with Y=v for which $X_i=u$

$$P(X_i = u|Y = v) = \frac{Count(X_i = u \wedge Y = v)}{Count(Y = v)}$$

Naïve Bayes (zero issue)

- In practice, some of these counts can be zero
- Fix this by adding "virtual" counts:

$$P(X_i = u|Y = v) = \frac{Count(X_i = u \wedge Y = v) + 1}{Count(Y = v) + 2}$$

  - (This is like putting a prior on parameters and doing MAP estimation instead of MLE)
  - This is called *Smoothing*

Underflow Prevention

- Multiplying lots of probabilities

➔ floating-point underflow.

- Recall: $\log(xy) = \log(x) + \log(y)$,

➔ better to sum logs of probabilities rather than multiplying probabilities.

Conditional Independence

- Probabilities of going to the beach and getting a heat stroke are not independent $P(B,S) > P(B)\ P(S)$
- May be independent, if we know the weather is hot $P(B,S|H) = P(B|H)\ P(S|H)$
- Hot weather explains the dependence between beach and heatstroke