Learning Distributions
Maximum Likelihood Estimate (MLE)
Bayes Classifier

Aarti Singh

Machine Learning 10-315
Sept 9, 2020

Logistics

- Anonymous feedback form

- Recitation on Friday Sept 11 – MLE/MAP + Optimization methods review and hands-on exercises

- QnA1 due TODAY

- HW1 to be released TODAY

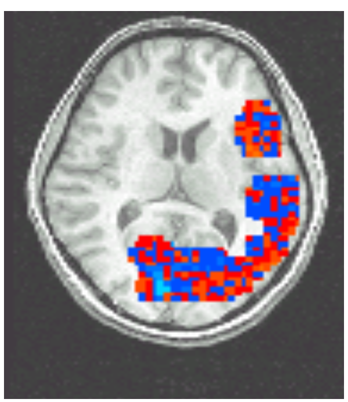
Why is ML not …

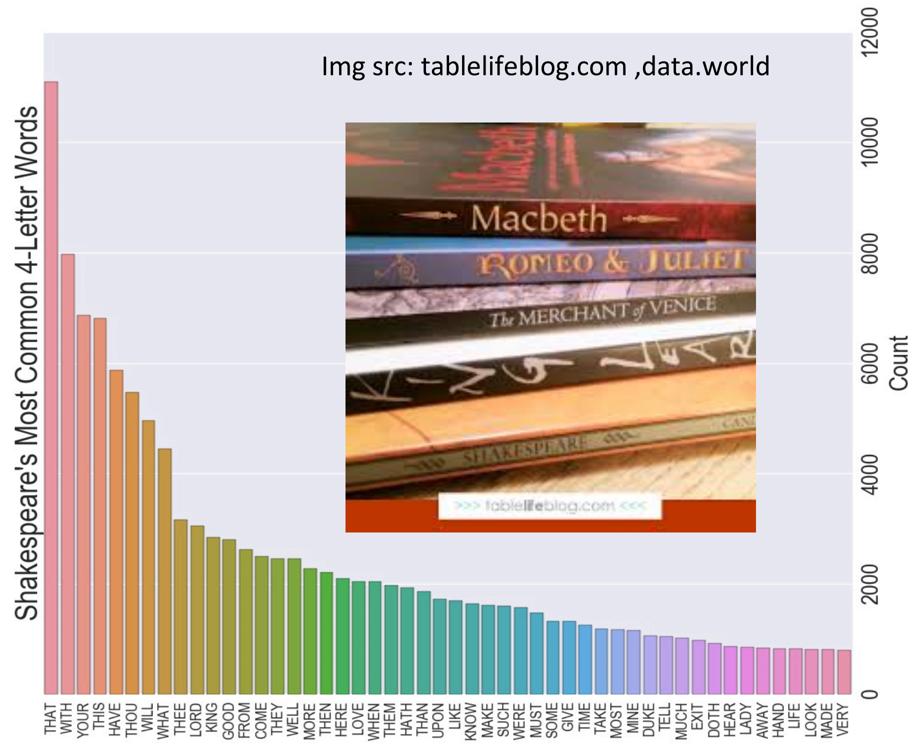
- Interpolation?
  - Noise, stochasticity, transfer across domains, …
- Statistics?
  - care about computationally efficiency (feasible, at least polynomial time in input size but typically much faster)
- Optimization?
  - Don't know true objective function, only stochastic version computed using data samples
- Data mining?
  - Generalization on new unseen data
- Your question?

Unsupervised Learning

Learning a Distribution

Bias of a coin





Distribution of words in text

Distribution of brain activity under stimuli

Notion of "Features aka Attributes"

Input $X \in \mathcal{X}$

Document/Article

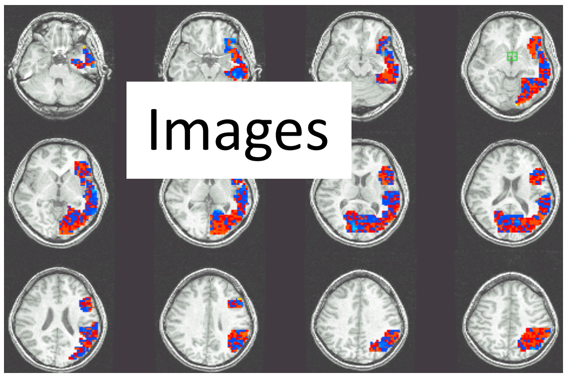
remember to wake up when class ends
=
wake ends to class remember up when

How to represent inputs mathematically?

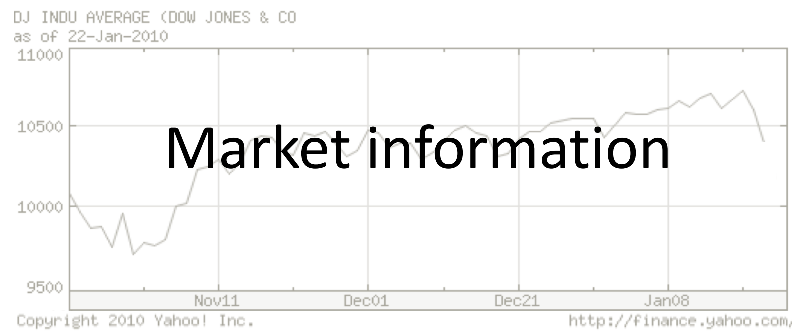
- Document vector X
  - ➢ Ideas?
  - list of words (different length for each document)
  - frequency of words (length of each document = size of vocabulary), also known as **Bag-of-words** approach
  - ➢ Why might this be limited?
  - Misses out context!!
  - list of n-grams (n-tuples of words)

Notion of “Features aka Attributes”

Input $X \in \mathcal{X}$



Input $X \in \mathcal{X}$



How to represent inputs mathematically?

- Image X = intensity/value at each pixel, fourier transform values, SIFT etc.
- Market information X = daily/monthly? price of share for past 10 years

Distribution of Inputs

Input $X \in \mathcal{X}$

Discrete Probability Distribution P(X) = P(X=x)

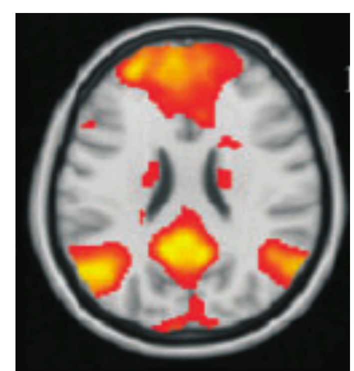
e.g. P(head) = ½, P(word x in text) = p_x

Probabilities in a distribution sum to 1

$\sum_x P(X=x) = 1$ P(tail) = 1 – p(head), $\sum_x p_x = 1$

Continuous Probability density p(x) P(a<=X<=b) = $\int_a^b p(x)dx$

e.g. p(brain activity)



Probability density integrate to 1

$$\int p(x)dx = 1$$

Distributions in Supervised tasks

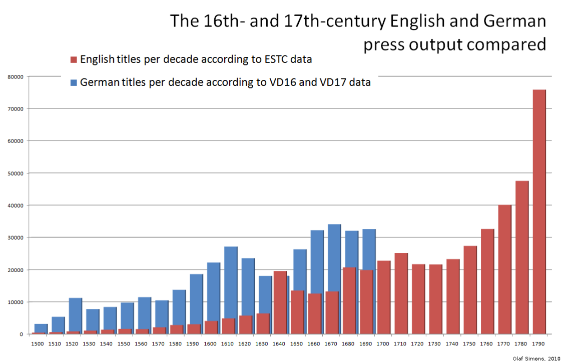
Input $X \in \mathcal{X}$

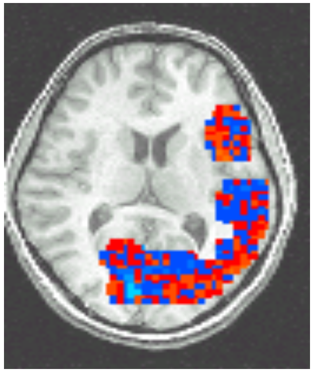
- Distribution learning also arises in supervised learning tasks e.g. classification

P(Y= y) Distribution of class labels

P(X = x | Y = y) Distribution of words in 'news' documents

Distribution of brain activity under 'stress'



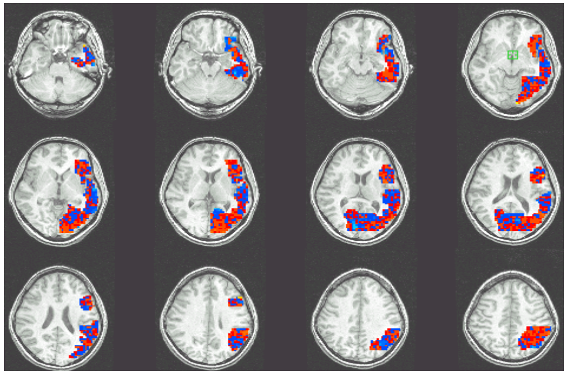


Olaf simons'10

P(Y = y|X = x) Distribution of topics given document

Classification

<u>Goal</u>: Construct **prediction rule** $f : \mathcal{X} \rightarrow \mathcal{Y}$



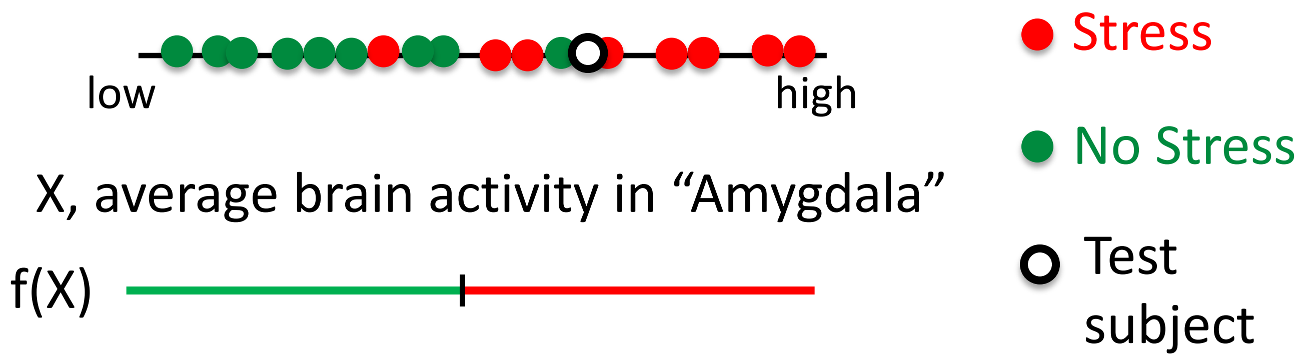
High Stress
Moderate Stress
Low Stress

Input feature vector, X **Label, Y**

In general: label Y can belong to more than two classes
X is multi-dimensional (many features represent an input)

But lets start with a simple case:
label Y is binary (either “Stress” or “No Stress”)
X is average brain activity in the “Amygdala”

Binary Classification



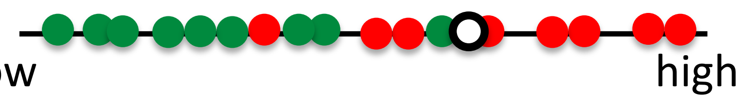
Model X and Y as random variables with joint distribution P_{XY}

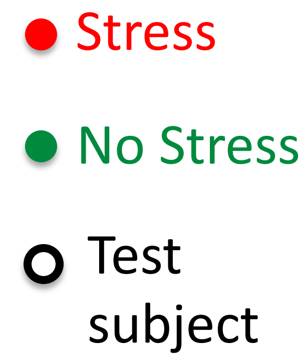
Training data $\{X_i, Y_i\}^n_{i=1}$ ~ iid (<u>independent</u> and <u>identically distributed</u>) samples from P_{XY}

Test data {X,Y} ~ iid sample from P_{XY}

Training and test data are independent draws from **<u>same</u>** distribution

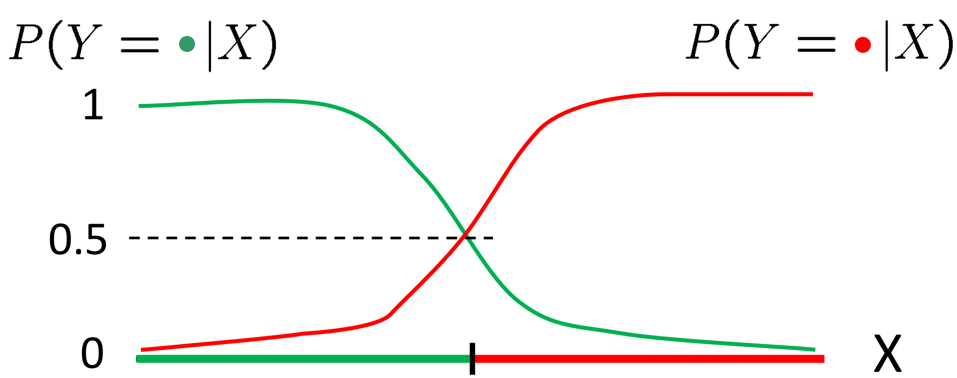
Bayes Classifier





X, average brain activity in "Amygdala"

Model X and Y as random variables



For a given X, f(X) = label Y which is more likely

$$f(X) = \arg\max_{Y=y} P(Y = y|X = x)$$

Bayes Rule

Bayes Rule: $$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

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

To see this, recall:

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

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

Thomas Bayes

Bayes Classifier

Bayes Rule:

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

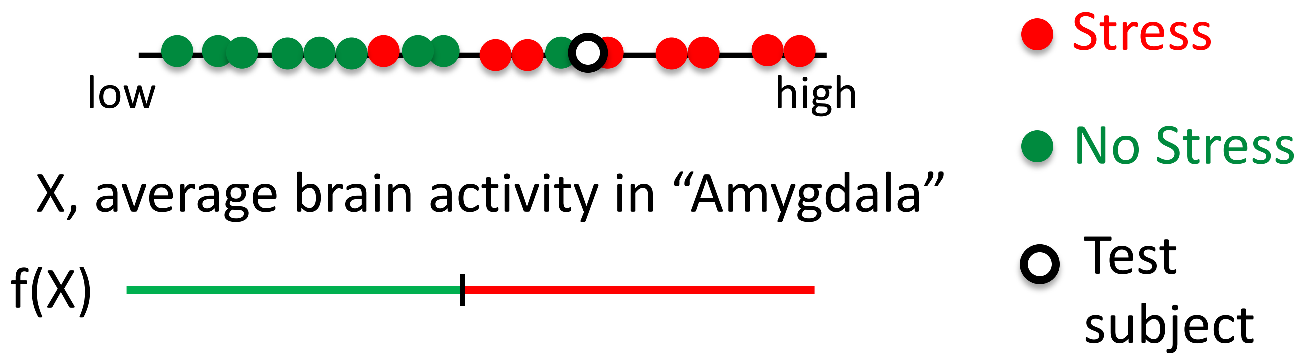
$$P(Y = y|X = x) \;=\; \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

Bayes classifier:

$$f(X) = \arg\max_{Y=y} P(Y = y|X = x)$$

$$= \arg\max_{Y=y} \underbrace{P(X = x|Y = y)}_{\text{Class conditional Distribution of features}}\underbrace{P(Y = y)}_{\text{Distribution of class}}$$

Bayes Classifier



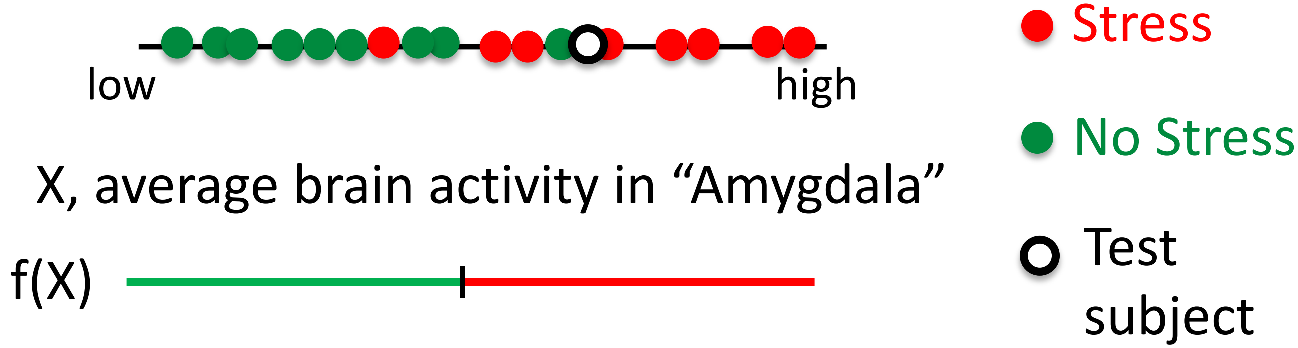
$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

We can now consider appropriate distribution models for the two terms:

Class distribution P(Y=y)

Class conditional distribution of features P(X=x|Y=y)

Modeling class distribution



Modeling Class distribution P(Y=y) = Bernoulli(θ)

$$P(Y = \bullet) = \theta \qquad\qquad P(Y = \bullet) = 1 - \theta$$

Like a coin flip

How to learn parameters from data? MLE

(Discrete case)

Learning parameters in distributions

$$P(Y = {\color{red}\bullet}) = \theta \qquad\qquad P(Y = {\color{green}\bullet}) = 1 - \theta$$

Learning θ is equivalent to learning probability of head in coin flip.

- How do you learn that?

Data =

Answer: 3/5

- Why??

Bernoulli distribution

Data, D = 

- P(Heads) = θ, P(Tails) = 1-θ
- Flips are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Bernoulli distribution

<u>Choose θ that maximizes the probability of observed data aka Likelihood</u>

Maximum Likelihood Estimation (MLE)

Choose θ that maximizes the probability of observed data (aka likelihood)

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

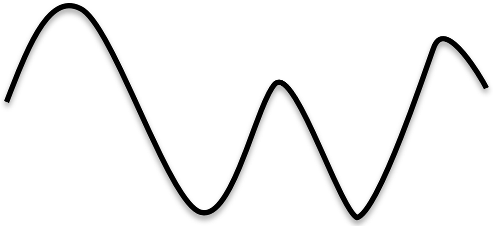
MLE of probability of head:

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$ **= 3/5**

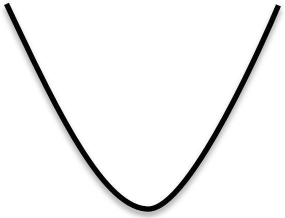
"Frequency of heads"

Short detour - Optimization

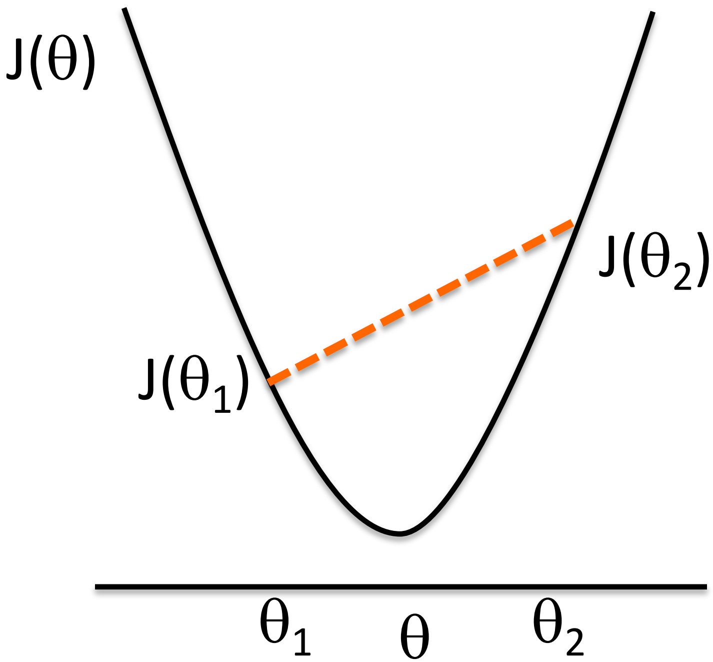
- Optimization objective $J(\theta)$
- Minimum value $J^* = \min_\theta J(\theta)$
- Minima (points at which minimum value is achieved) may not be unique



- If function is strictly convex, then minimum is unique

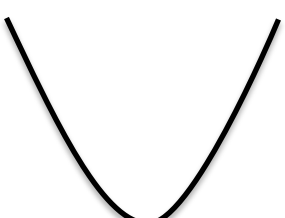


Convex functions

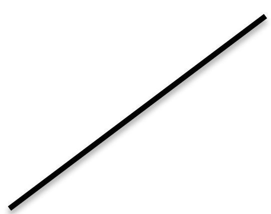


A function $J(\theta)$ is called **convex** if the line joining two points $J(\theta_1), J(\theta_2)$ on the function does not go below the function on the interval $[\theta_1, \theta_2]$

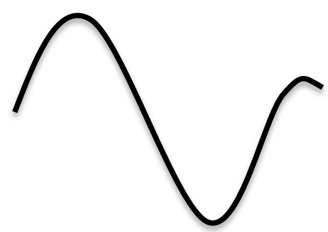
(Strictly) Convex functions have a unique minimum!



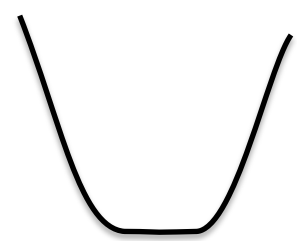
Convex



Both Concave & Convex



Neither



Convex but not strictly convex

Optimizing convex (concave) functions

- Derivative of a function
  - Partial derivative
- Derivative is zero at minimum of a convex function
- Second derivative is positive at minimum of a convex function

Optimizing convex (concave) functions

- What about

  concave functions?

  non-convex/non-concave functions?

  functions that are not differentiable?

  optimizing a function over a bounded domain aka constrained optimization?

Maximum Likelihood Estimation (MLE)

Choose θ that maximizes the probability of observed data (aka likelihood)

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

MLE of probability of head:

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

= 3/5

"Frequency of heads"

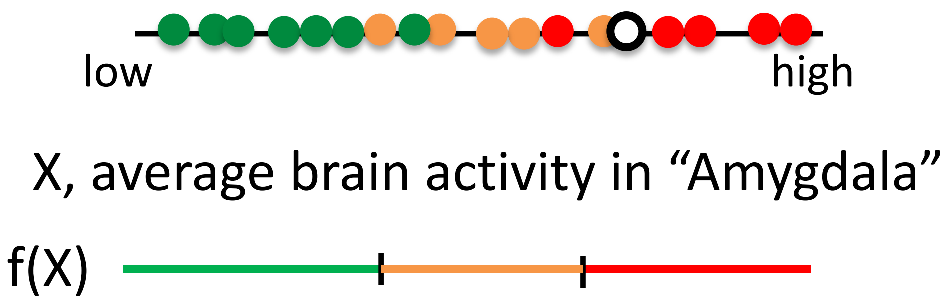
Derivation

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

Derivation

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

Modeling class distribution



- ● High Stress
- ● Moderate Stress
- ● Low Stress
- ○ Test subject

➢ How do we model multiple (>2) classes?

Modeling Class distribution P(Y) = Multinomial(p_H,p_M,p_L)

$$P(Y = \bullet) = p_H \quad P(Y = \bullet) = p_M \quad P(Y = \bullet) = p_L$$

Like a dice roll

$$p_H + p_M + p_L = 1$$

Multinomial distribution

Data, D = rolls of a dice

- $P(1) = p_1$, $P(2) = p_2$, …, $P(6) = p_6$ $p_1+\ldots+p_6 = 1$
- Rolls are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Multinomial(θ) distribution where

$$\theta = \{p_1, p_2, \ldots, p_6\}$$

Choose θ that maximizes the probability of observed data aka "Likelihood"

Maximum Likelihood Estimation (MLE)

Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} \; P(D \mid \theta)$$

MLE of probability of rolls:

$$\widehat{\theta}_{MLE} = \hat{p}_{1,MLE}, \ldots, \hat{p}_{6,MLE}$$

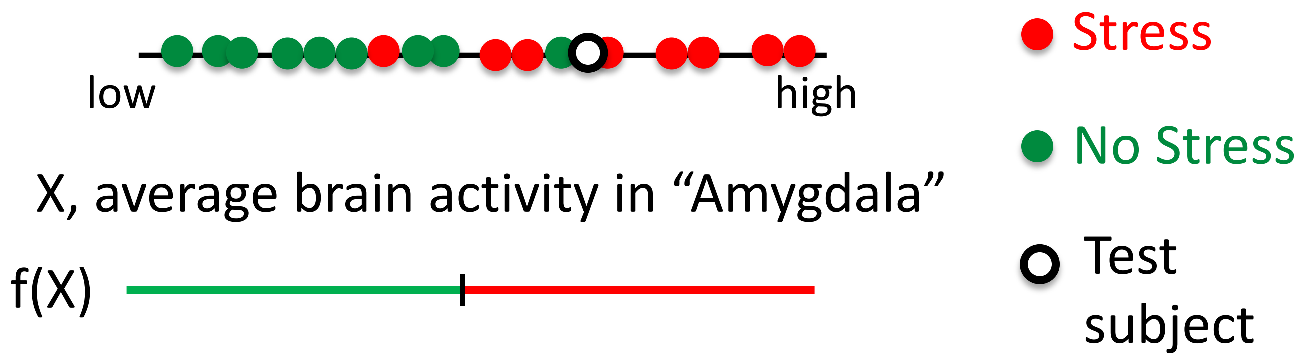
$$\hat{p}_{y,MLE} = \frac{\alpha_y}{\sum_y \alpha_y}$$

Rolls that turn up y

Total number of rolls

"Frequency of roll y"

Bayes Classifier



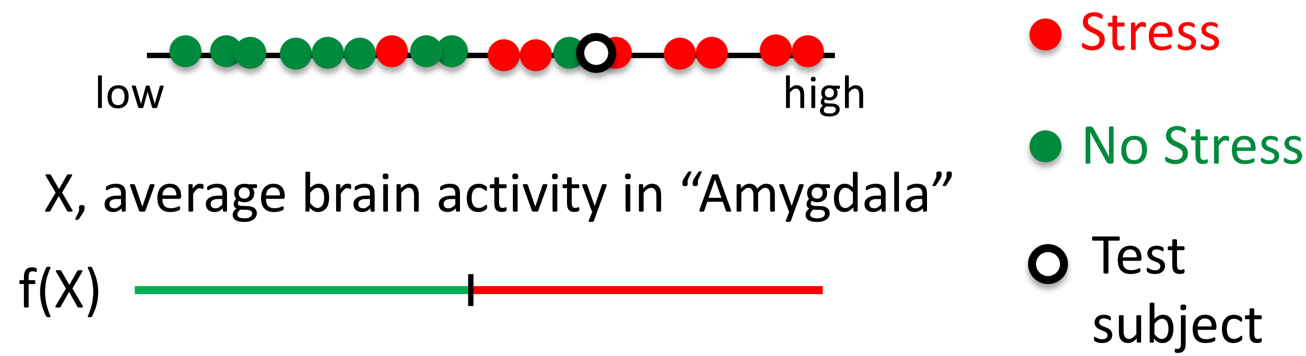
$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

We can now consider appropriate distribution models for the two terms:

Class distribution P(Y=y)

Class conditional distribution of features P(X=x|Y=y)

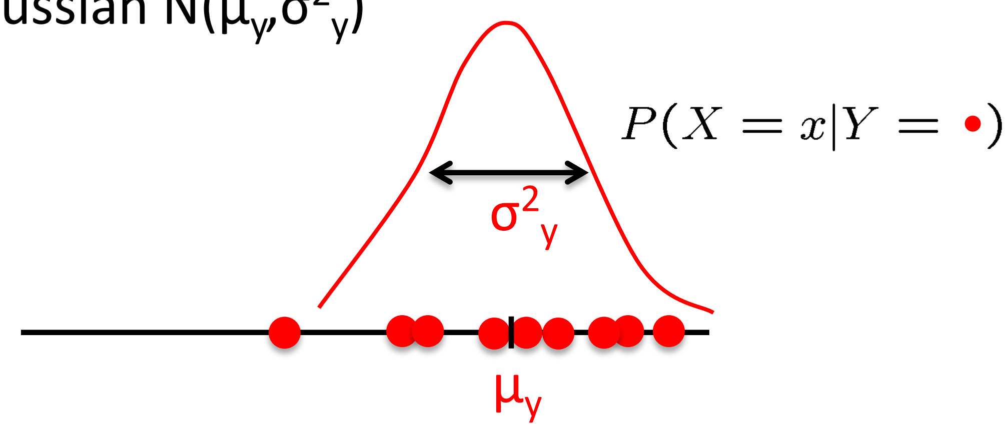
Modeling class conditional distribution of features



Modeling class conditional distribution of feature P(X=x|Y=y)

➢ What distribution would you use?

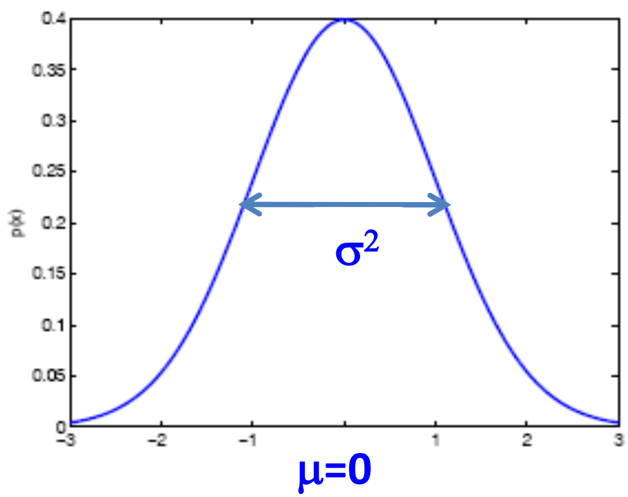
E.g. P(X=x|Y=y) = Gaussian $N(\mu_y,\sigma^2_y)$

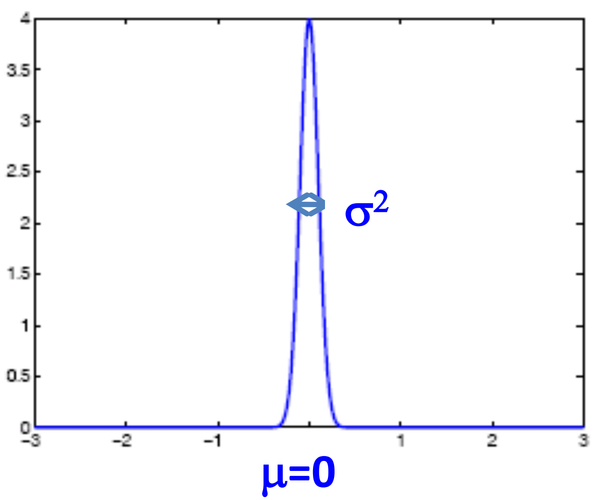


1-dim Gaussian distribution

X is Gaussian $N(\mu,\sigma^2)$

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





Why Gaussian?

- Properties
  - Fully Specified by first and second order statistics
    - Uncorrelated ⇔ Independence
  - X, Y Gaussian => aX+bY Gaussian

  - <u>Central limit theorem:</u> if $X_1, \ldots, X_n$ are any iid random variables with mean μ and variance $\sigma^2 < \infty$

    then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \sim N(0, \sigma^2)$$

d-dim Gaussian distribution

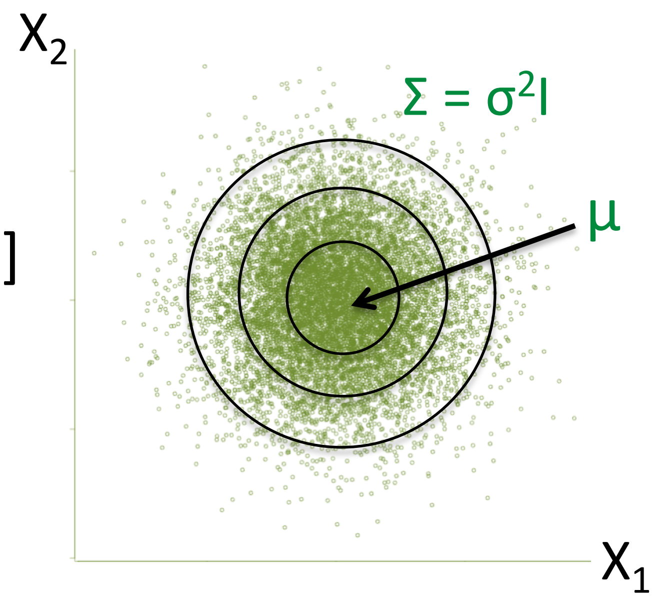
X is Gaussian N(μ, Σ)

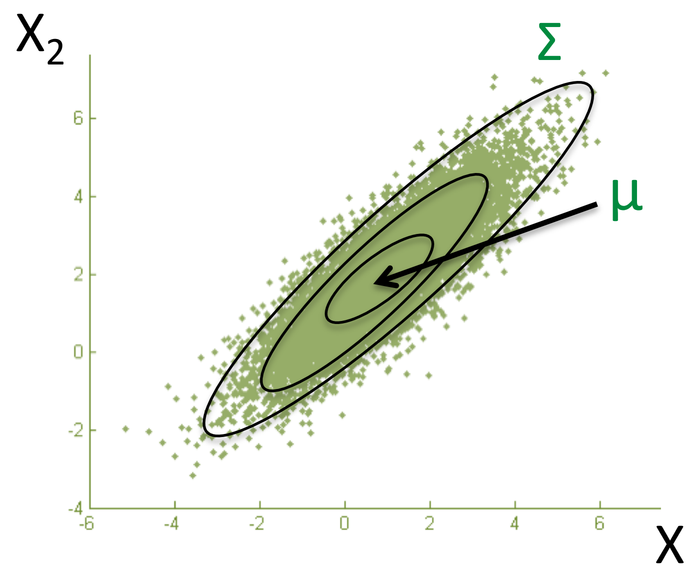
μ is d-dim vector, Σ is dxd dim matrix

$$P(X = x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

d=2

$X = [X_1; X_2]$





How to learn parameters from data? MLE

(Continuous case)

Gaussian distribution

Data, D =

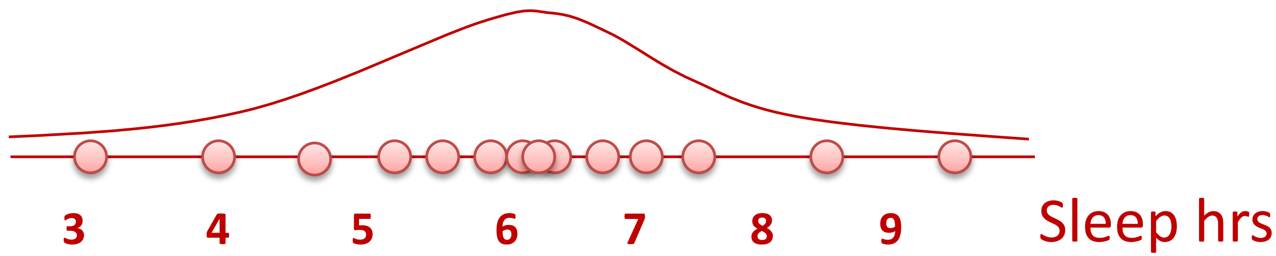
3 **4** **5** **6** **7** **8** **9** Sleep hrs

How many hours did you sleep last night?

- Poll

Gaussian distribution

Data, D =



- Parameters: μ – mean, σ^2 - variance
- Sleep hrs are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Gaussian distribution

Maximum Likelihood Estimation (MLE)

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

Maximum Likelihood Estimation (MLE)

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X_i-\mu)^2/2\sigma^2}$$

Identically distributed

Maximum Likelihood Estimation (MLE)

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X_i-\mu)^2/2\sigma^2}$$

Identically distributed

$$= \arg \max_{\theta=(\mu,\sigma^2)} \underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}}_{J(\theta)}$$

Derivation

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

- Breakout

Groups 1-10: Jamboard_1_10
Groups 11-20: Jamboard_11_20

MLE for Gaussian mean and variance

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

Gaussian Bayes classifier

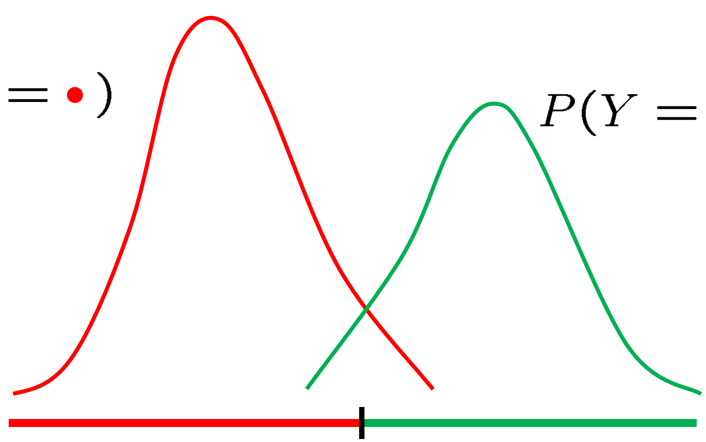
$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

How to learn parameters θ, μ_y, Σ_y from data?

Gaussian(μ_y, Σ_y)

Bernoulli(θ)

$P(Y=\bullet)P(X=x|Y=\bullet)$



$P(Y=\bullet)P(X=x|Y=\bullet)$

1-dim Gaussian Bayes classifier

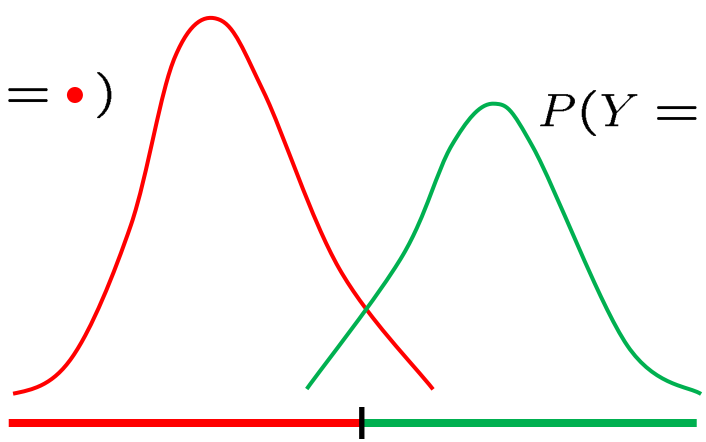
$$f(X) = \arg\max_{Y=y} \underbrace{P(X = x|Y = y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y = y)}_{\text{Class distribution}}$$

Gaussian(μ_y, σ^2_y)

Bernoulli(θ)

- What decision boundaries can we get in 1-dim?

$P(Y = \bullet)P(X = x|Y = \bullet)$



$P(Y = \bullet)P(X = x|Y = \bullet)$

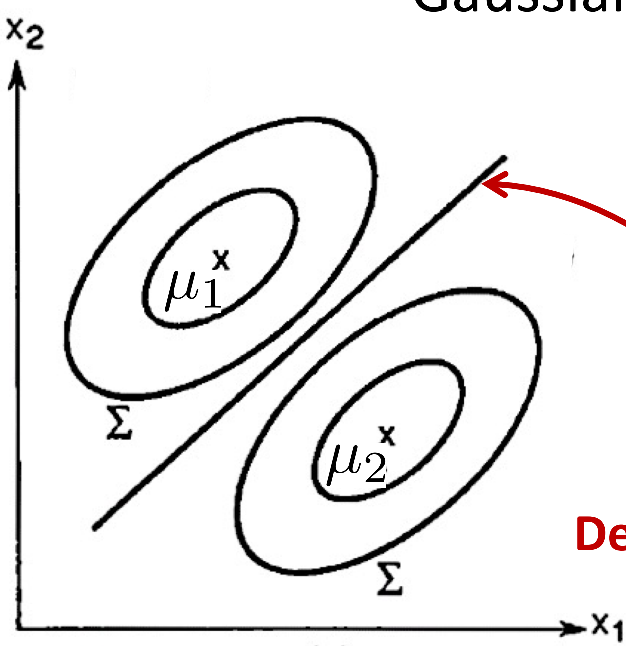
d-dim Gaussian Bayes classifier

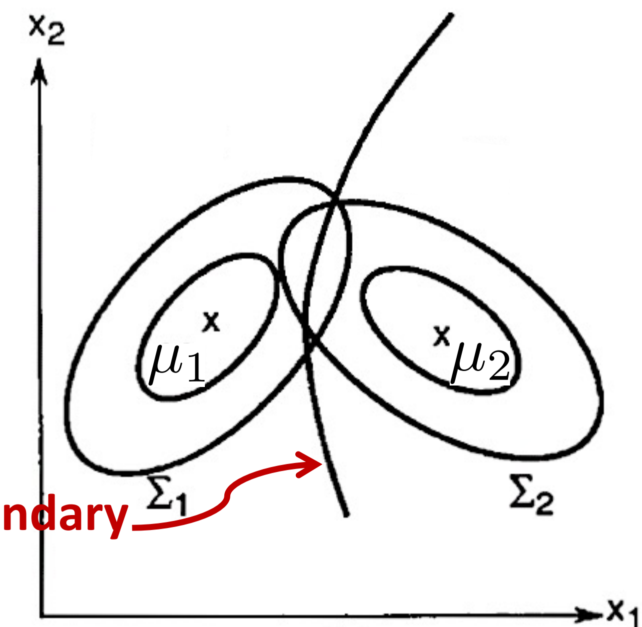
$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

- What decision boundaries can we get in d-dim?

Class conditional Distribution of features → Gaussian(μ_y,Σ_y)

Class distribution → Bernoulli(θ)





Decision Boundary

Decision Boundary of Gaussian Bayes

- Decision boundary is set of points x: P(Y=1|X=x) = P(Y=0|X=x)
- By Bayes theorem, equivalent to x:

<u>Lets find the decision boundary.</u>

If class distribution is P(Y=1) = Ber(θ) and

class conditional feature distribution P(X=x|Y=y) is 2-dim Gaussian N(μ_y,Σ_y)

$$P(X=x|Y=y) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_y|}} \exp\left(-\frac{(x-\mu_y)\Sigma_y^{-1}(x-\mu_y)'}{2}\right)$$

Decision Boundary of Gaussian Bayes

- Decision boundary is set of points x: P(Y=1|X=x) = P(Y=0|X=x)

Compute the ratio

$$1 = \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \frac{P(X=x|Y=1)P(Y=1)}{P(X=x|Y=0)P(Y=0)}$$

$$= \sqrt{\frac{|\Sigma_0|}{|\Sigma_1|}} \exp\left(-\frac{(x-\mu_1)\Sigma_1^{-1}(x-\mu_1)'}{2} + \frac{(x-\mu_0)\Sigma_0^{-1}(x-\mu_0)'}{2}\right) \frac{\theta}{1-\theta}$$

In general, this implies a quadratic equation in x. But if $\Sigma_1 = \Sigma_0$, then quadratic part cancels out and decision boundary is linear.

d-dim Gaussian Bayes classifier

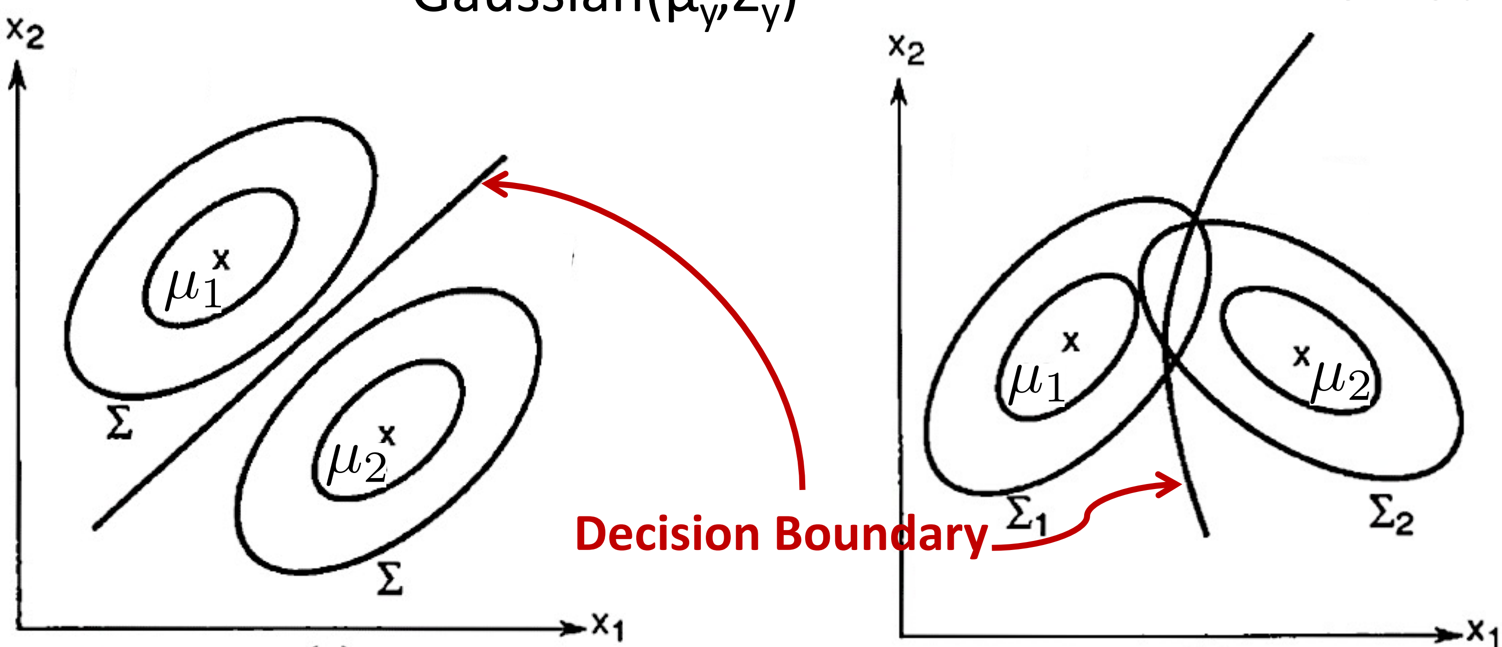
$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

Distribution of features

Gaussian(μ_y,Σ_y)

Bernoulli(θ)

- What decision boundaries can we get in d-dim?



Glossary of Machine Learning

- Feature/Attribute
- iid
- Bayes classifier
- Class distribution
- Class conditional distribution of features
- Estimator – hat notation
- MLE
- Decision boundary