Bayes classifier, Decision boundary

Aarti Singh

Machine Learning 10-315
Sept 1, 2021

In-person Expectations

- Please **do not come to class if you feel sick** or have symptoms of a potentially contagious illness.

- This holds <u>even if you know you do not have COVID-19</u> but have symptoms that may be a sign of other contagious illnesses, such as a <u>cold or flu</u>.

- Coming to class when sick is NOT a sign of hard work. Be responsible, care for your classmates and campus community.

Announcements

- Canvas fixed
- Late days – total 4, no more than 1 for a QnA, no more than 2 for a HW
- Tentative HW due dates have been posted
- QnA1 out today
  – due in 1 week, Sept 8 11:59 pm ET
- Office hours

Day	Time	Location	Staff
Mondays	1:30 pm	TBA	Christina
Tuesdays	3:30 pm	TBA	Spoorthi
Wednesdays	2:30 pm	TBA	Haitian
Thursdays	10:30 am	Zoom (link on Canvas)	Aarti

Notion of "Features aka Attributes"

Input $X \in \mathcal{X}$

remember to wake up when class ends

=

wake ends to class remember up when

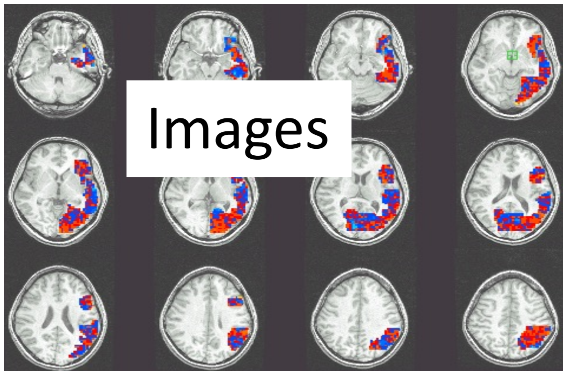
How to represent inputs mathematically?

- 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?

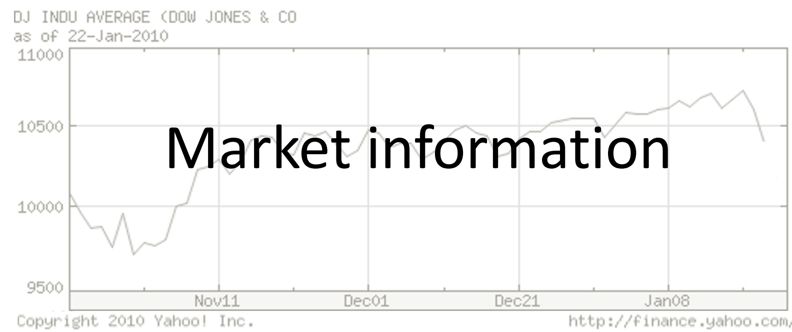
    Misses out context!!
  - list of n-grams (n-tuples of words)

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)

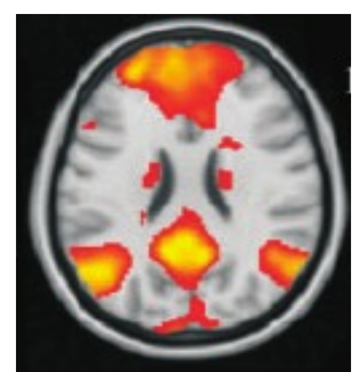
e.g. P(head) = ½, P(word x in text) = p_x

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

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

$P(X,Y)$

Distributions in Supervised tasks

$p(X)$ $p(Y)$ $p(X|Y=yes)$

$P(Y|X=x)$

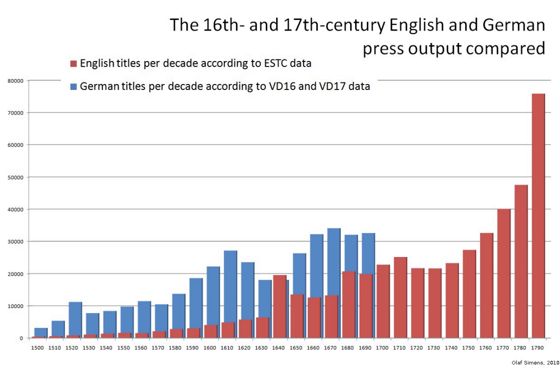
Input $X \in \mathcal{X}$

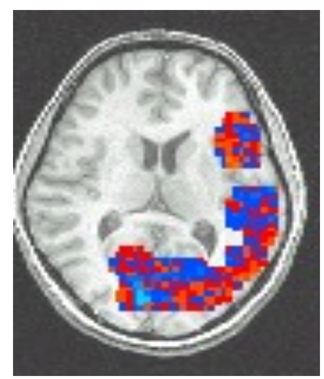
- Distribution learning also arises in supervised learning tasks e.g. classification

P(Y= y) Distribution of class labels

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

Distribution of brain activity under ‘stress’





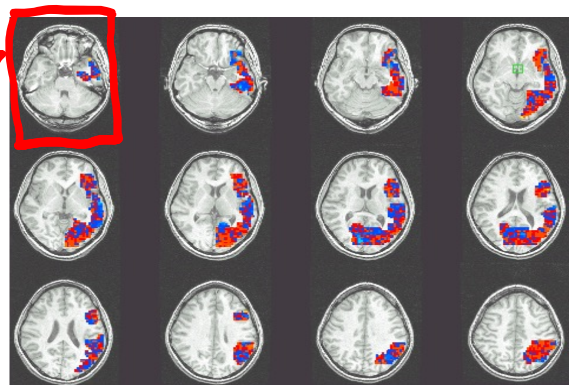
Olaf simons’10

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

Classification

Goal: Construct **prediction rule** $f \ : \mathcal{X} \rightarrow \mathcal{Y}$

$x=$



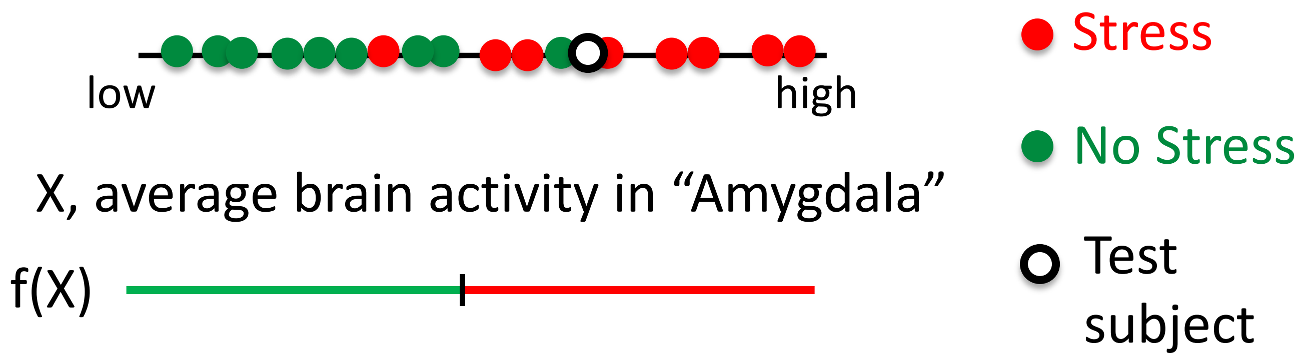
High Stress
Moderate Stress
Low Stress

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



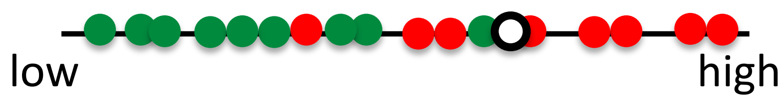
Model X and Y as random variables with joint distribution P_{XY}

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

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

Training and test data are independent draws from **same** distribution

Bayes Classifier



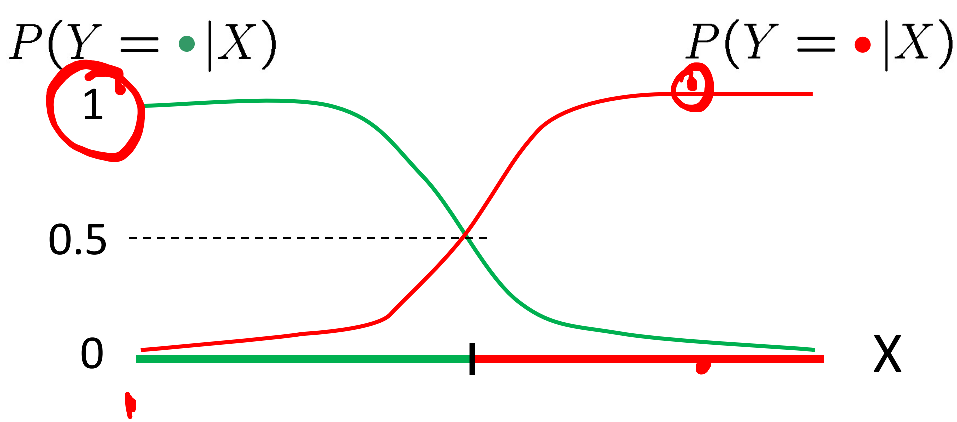
● Stress

● No Stress

○ Test subject

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

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

$f: X \rightarrow Y$

Optimality of Bayes Classifier

$$\min_f P(f(X) \neq Y) \longrightarrow E[1_{f(X)\neq Y}]$$

$$P(A) = \sum_b P(A,B) = \sum_b P(A|B)P(B)$$

$$P(f(X) \neq Y) = \int \underbrace{P(f(X) \neq Y \mid X = x)\, P(X = x)}_{P(f(X)\neq Y,\, X=x)}\, dx$$

$$= \int_{x: f(x)=0} P(Y=1 \mid X=x)\, P(X=x)\, dx + \int_{x: f(x)=1} P(Y=0 \mid X=x) P(X=x)\, dx$$

This is minimized if

$$f(x) = \begin{cases} 0 & P(Y=1 \mid X=x) < P(Y=0 \mid X=x) \\ 1 & > \end{cases}$$

$$= \arg\max_y P(Y=y \mid X=x)$$

Bayes Rule

$P(A,B) = P(A|B)P(B)$
$= P(B|A)P(A)$

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 – equivalent form

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

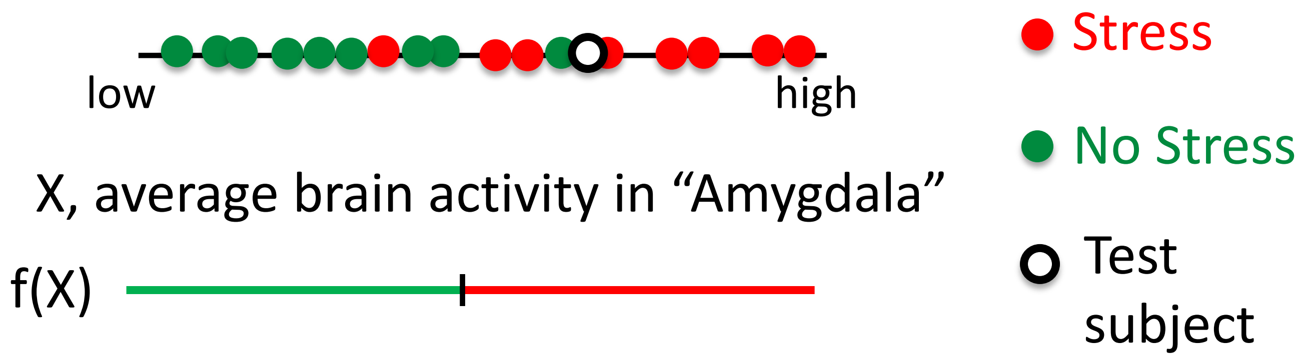
$$P(Y = y|X = x) \;=\; \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

Bayes classifier:

$$\begin{aligned} f(X) = \; & \arg\max_{Y=y} \boxed{P(Y = y|X = x)} \\ = \; & \arg\max_{Y=y} \underbrace{P(X = x|Y = y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y = y)}_{\text{Distribution of class}} \end{aligned}$$

X=

Bayes Classifier



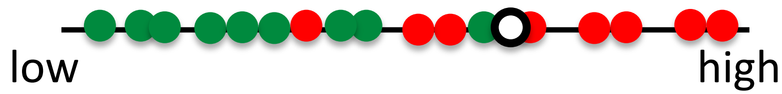
$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

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



● Stress

● No Stress

○ Test subject

X, average brain activity in “Amygdala”

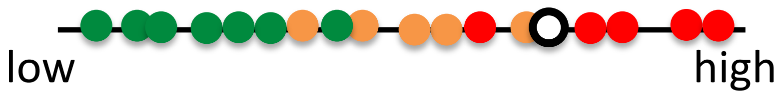
f(X)

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

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

Like a coin flip

Modeling class distribution



- High Stress
- Moderate Stress
- Low Stress
- Test subject

X, average brain activity in “Amygdala”

f(X) 

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

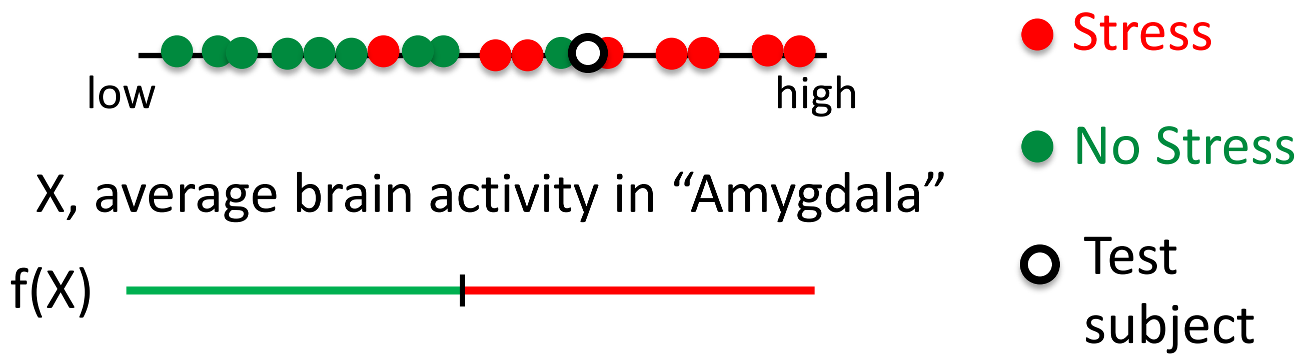
Modeling Class distribution P(Y) = Multinomial(p_H,p_M,p_L)

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

Like a dice roll

$$p_H + p_M + p_L = 1$$

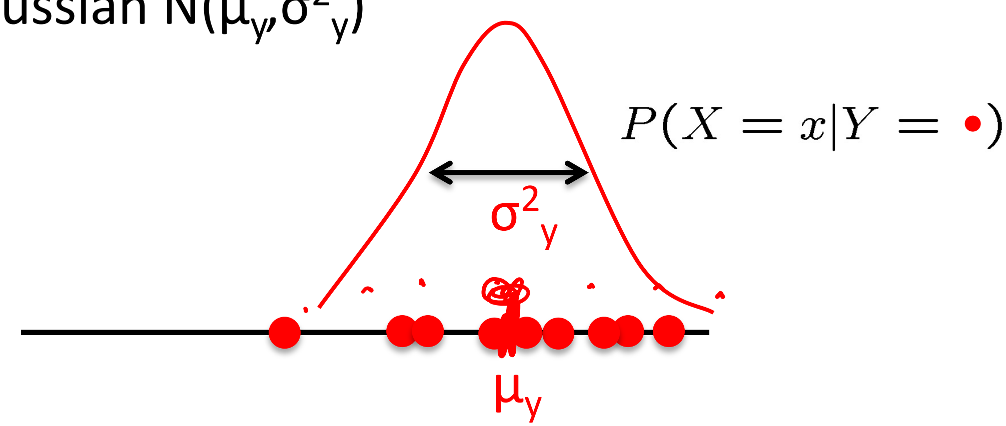
Modeling class conditional distribution of features



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

➢ What distribution would you use?

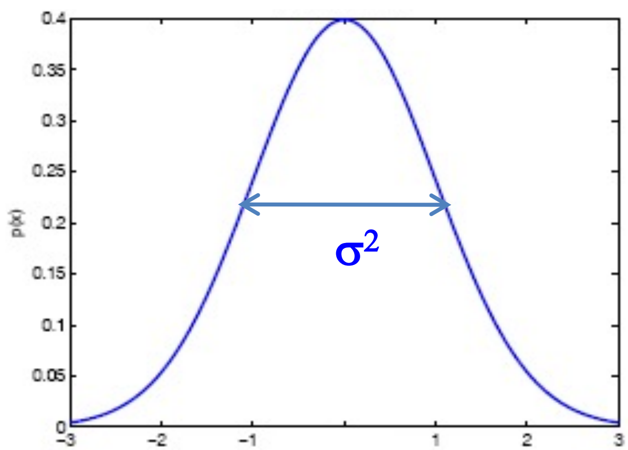
E.g. $P(X=x|Y=y)$ = Gaussian $N(\mu_y,\sigma^2_y)$



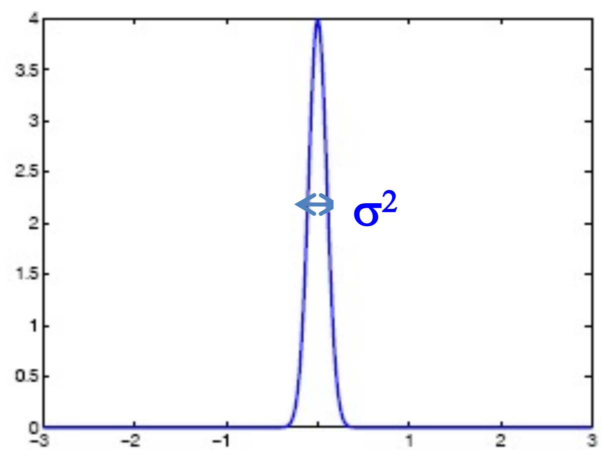
1-dim Gaussian distribution

X is Gaussian N(μ,σ²)

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



μ=0



μ=0

Why Gaussian?

$E[X^2]$

$\mu = E[X]$

$var(X) = \sigma^2$
$= E[(X - E[X])^2]$

- 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

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

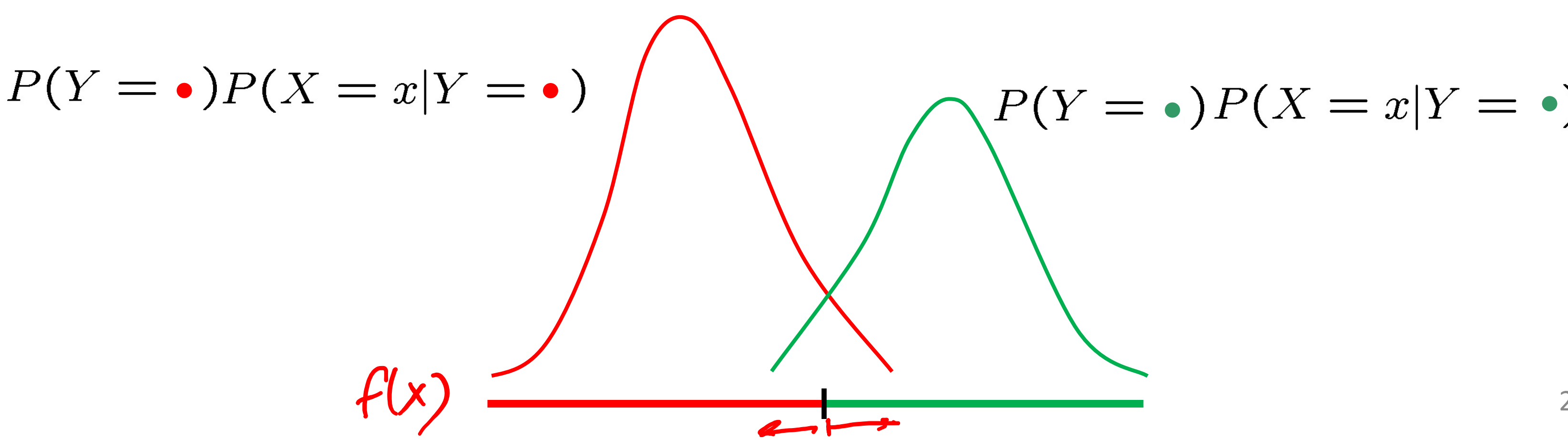
1-dim Gaussian Bayes classifier

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

Learn parameters θ, μ_y, σ_y from data

Class conditional Distribution of features $\rightarrow$ Gaussian(μ_y, σ^2_y)

Class distribution $\rightarrow$ Bernoulli(θ)



1-dim Gaussian Bayes classifier

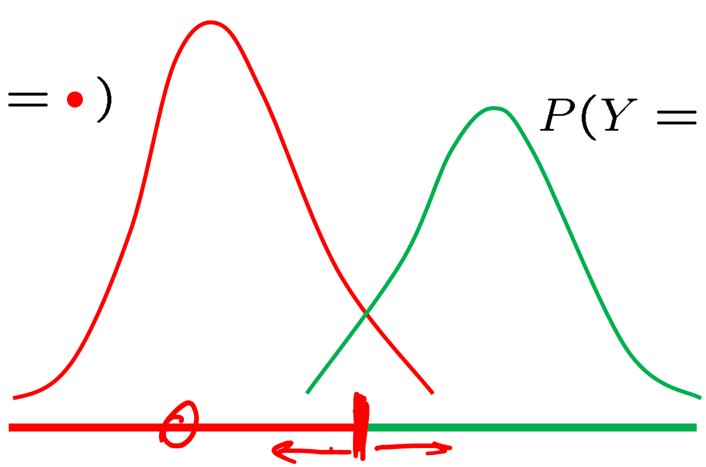
$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

Gaussian(μ_y, σ^2_y)

Bernoulli(θ)

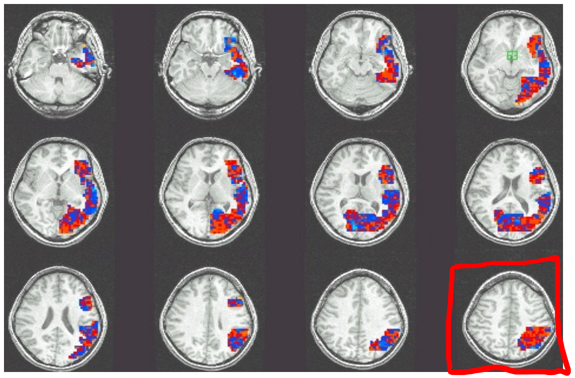
- What decision boundaries can we get in 1-dim?

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



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

d-dimensional inputs



High Stress
Moderate Stress
Low Stress

Input feature vector, X $= \begin{bmatrix} \vdots \end{bmatrix}_d$ **Label, Y**

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

➢ What distribution would you use?

$N(\mu, \sigma^2)$

1×1 1×1

E.g. P(X=x|Y=y) = Gaussian $N(\mu_y, \Sigma_y)$

d×1 d×d

$\Sigma_{ij} = E[(X_i - E[X_i])(X_j - E[X_j])]$ $\Sigma_{ii} = E[(X_i - E[X_i])^2]$

d-dim Gaussian distribution

X is Gaussian N(μ, Σ) μ is d-dim vector, Σ is dxd dim matrix

$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x-\mu)^2(\sigma^2)^{-1}}$

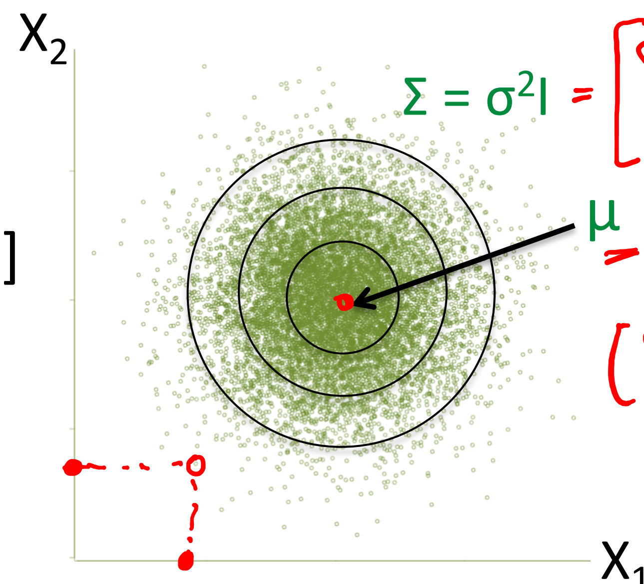
$$P(X = x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

dx1 dx1 dxd

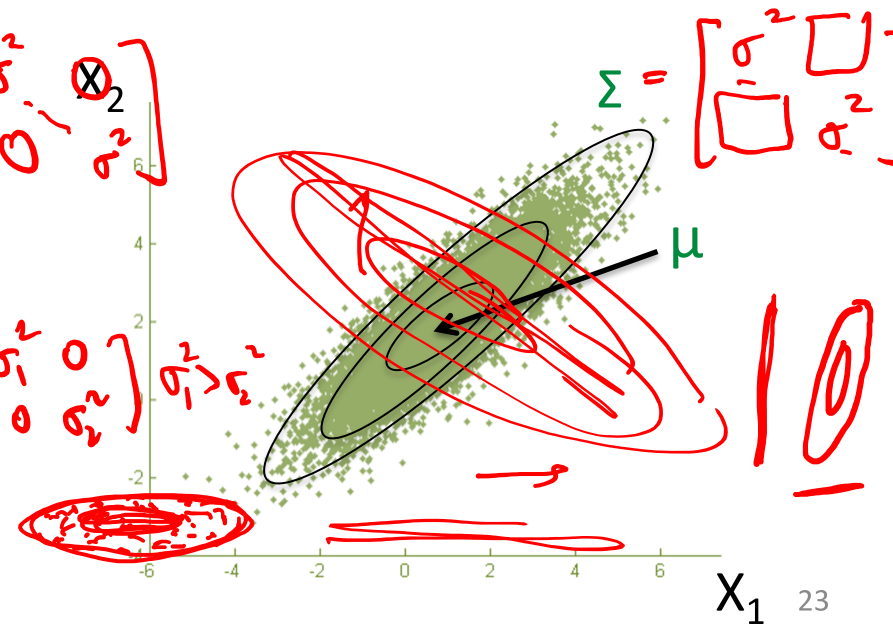
-0.5 ? 0.5

d=2

$X = [X_1; X_2]$



$\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$ $\sigma_1^2 > \sigma_2^2$



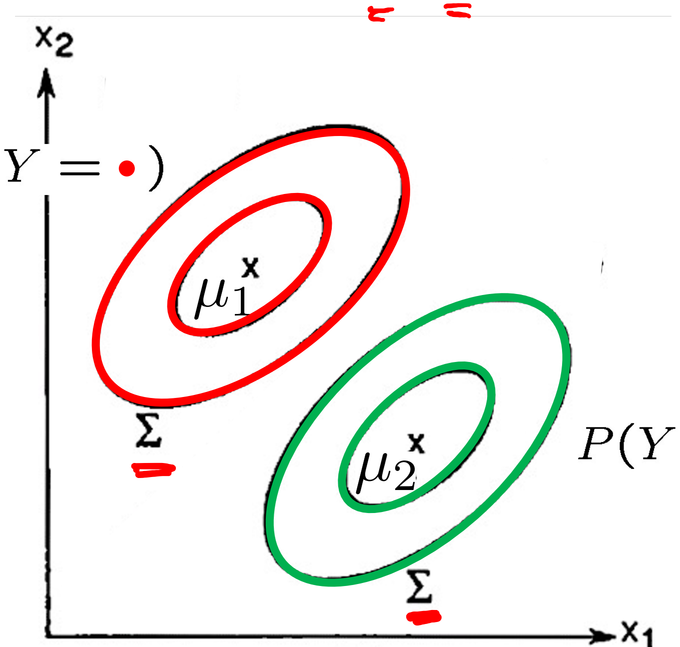
d-dim Gaussian Bayes classifier

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

Learn parameters θ, μ_y, Σ_y from data

Class conditional Distribution of features $\rightarrow$ Gaussian(μ_y, Σ_y)

Class distribution $\rightarrow$ Bernoulli(θ)



d-dim Gaussian Bayes classifier

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

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

Class distribution → Bernoulli(θ)

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