Bayes classifier, Decision boundary

Aarti Singh

Machine Learning 10-315
Jan 24, 2022

What is Machine Learning?

Design and Analysis of algorithms that

- improve their performance
- at some task
- with experience

(experience) **(performance on task)**

Tasks

Broad categories -

- **Supervised learning** $X \to Y$

  Classification, Regression

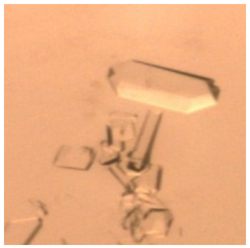
- **Unsupervised learning** X

  Density estimation, Clustering, Dimensionality reduction

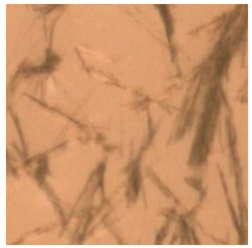
- Graphical models
- Semi-supervised learning
- Active learning
- Bayesian optimization
- Reinforcement learning
- Many more …

Experience

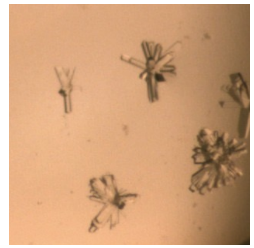
Training data



Crystal

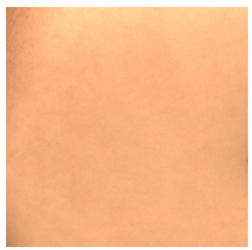


Needle

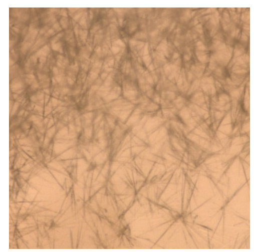


Tree

Tree

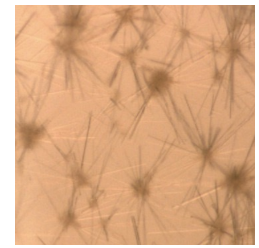


Empty



Needle

Test data



?

A good ML algorithm

should: **generalize** aka perform well on test data

should not: **overfit** the training data

gap between test & training performance

Performance

For test data X, measure of closeness between label *Y* and prediction *f*(*X*)

$f : X \to Y$

Binary Classification $\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}}$ 0/1 loss

Regression $\text{loss}(Y, f(X)) = (f(X) - Y)^2$ squared loss

$|f(x) - Y|$ abs loss

We will talk about more performance measures including for unsupervised learning later in course.

Training data $(x_i, y_i)_{i=1}^{n}$

$\frac{1}{n}\sum_{i=1}^{n} 1_{f(x_i) \neq y_i}$

Poll

- A classifier with 100% accuracy on training data and 70% accuracy on test data is better than a classifier with 80% accuracy on training data and 80% accuracy on test data.

  A. True B. False

- Which classifier is better, given following statistics on test accuracy?

	Mean	Best run	Std
Classifier A	92%	97%	15%
Classifier B	87%	100%	5%

Design of ML algorithms

x, y

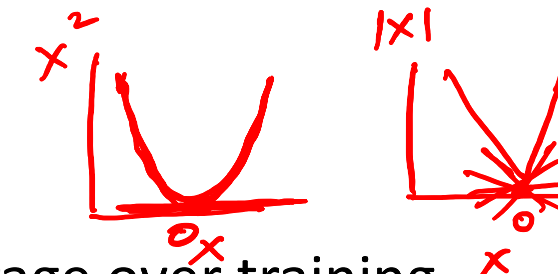
Minimize loss in expectation (over random test data)

$$\min_f E_{XY}[\mathrm{loss}(f(X),Y)] \qquad (f(x)-Y)^2$$

$\frac{\partial}{\partial f}$

- Different methods use
  - different loss, e.g. 0/1 loss, squared loss, etc.
  - different model f, e.g. linear, neural network, decision tree etc.

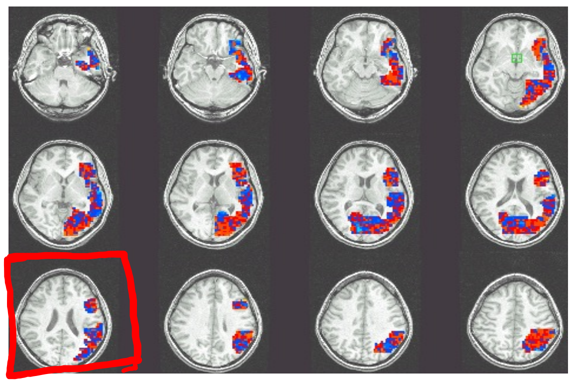
➢ Why prefer squared loss over abs loss?



- For training, replace expectation with average over training data

Classification

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



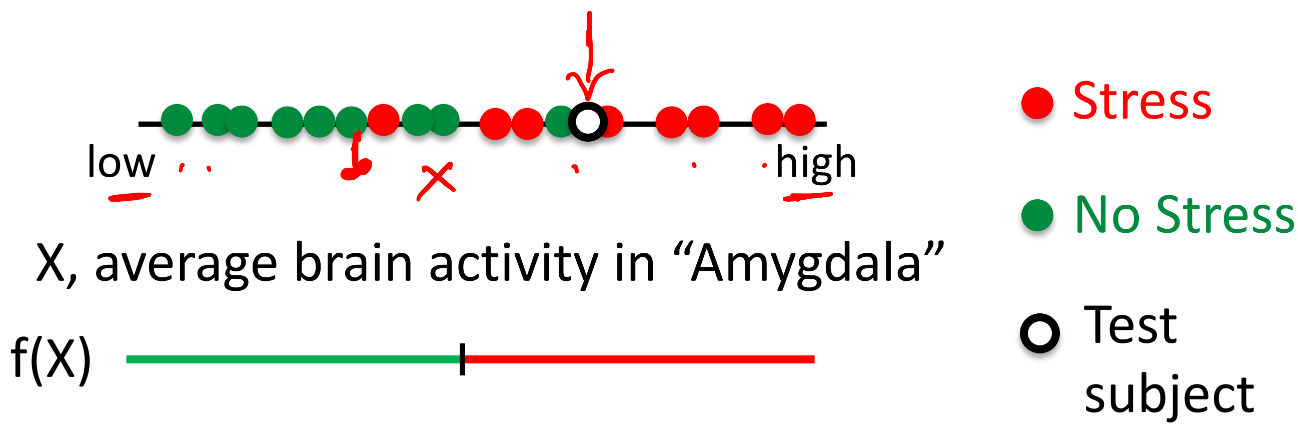
High Stress
Moderate Stress
Low Stress

Input, X **Label, Y**

In general: label Y can belong to more than two classes
X is multi-dimensional (brain activity in all regions)

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} unknown

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

Optimal classifier

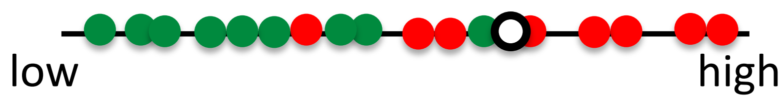
Minimize loss in expectation (over random test data)

$$\min_f E_{XY}[\text{loss}(f(X),Y)]$$

- Which classifier f is optimal for 0/1 loss, assuming we know data-generating distribution P(X,Y)?

$\{x_i, y_i\}_{i=1}^{n}$

Bayes Classifier



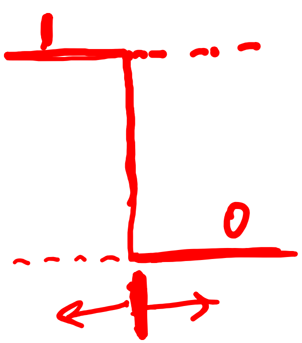
Stress

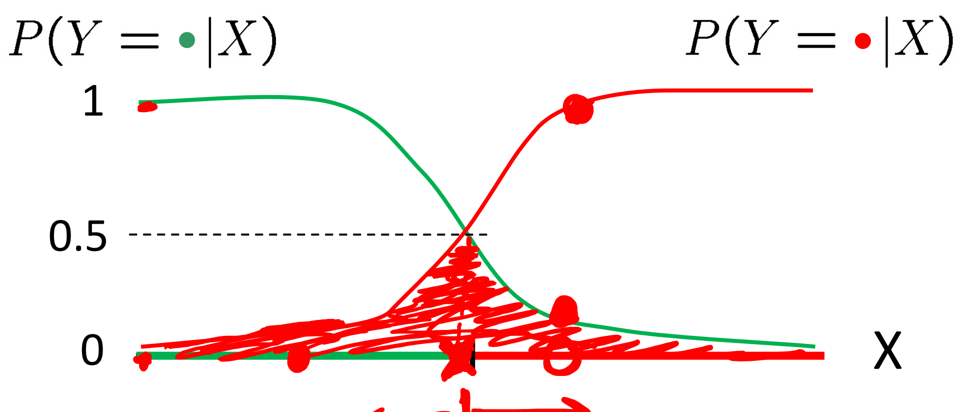
No Stress

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

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

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

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:

$\rightarrow$ P(X,Y) = P(X|Y) P(Y) $\leftarrow$ chain rule

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

Thomas Bayes

Bayes Classifier – equivalent form

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

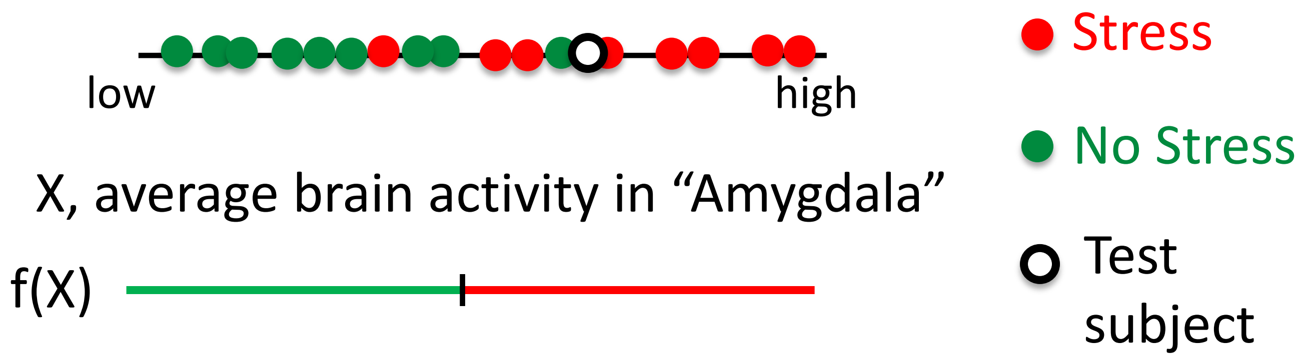
$$P(Y=y|X=x) = \frac{P(X=x|Y=y)P(Y=y)}{P(X=x)}$$

Bayes classifier:

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

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

Bayes Classifier

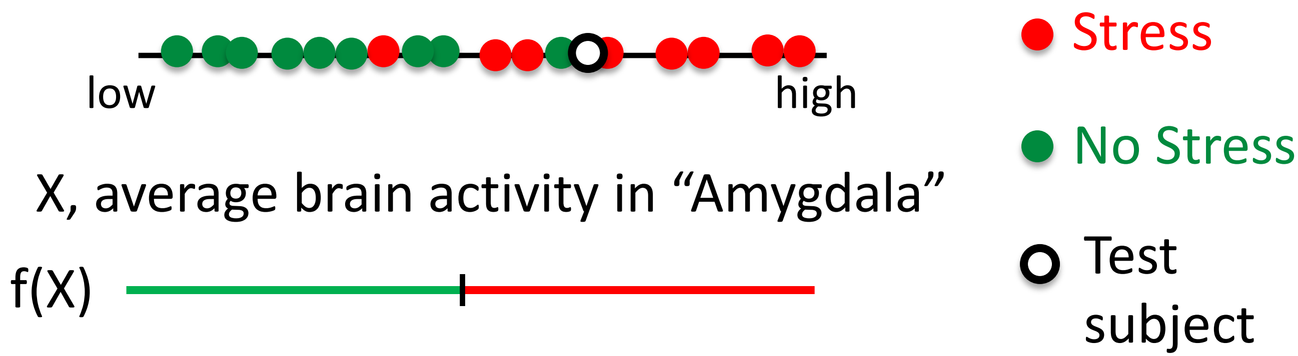


$$f(X) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\substack{\text{Class conditional}\\ \text{Distribution of inputs}}} \underbrace{P(Y=y)}_{\text{Class distribution}}$$

We can now consider appropriate models for the two terms:

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

Modeling class distribution

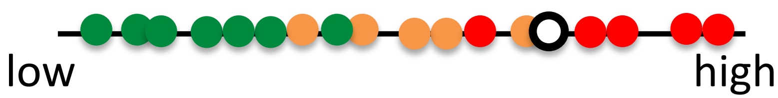


Modeling Class distribution P(Y=y) = Bernoulli(θ)

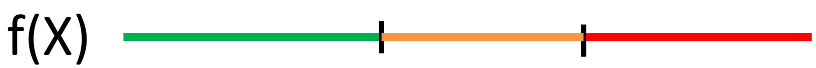
$$P(Y = {\color{red}\bullet}) = \theta \qquad\qquad P(Y = {\color{green}\bullet}) = 1 - \theta$$

Like a coin flip

Modeling class distribution



X, average brain activity in “Amygdala”



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

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

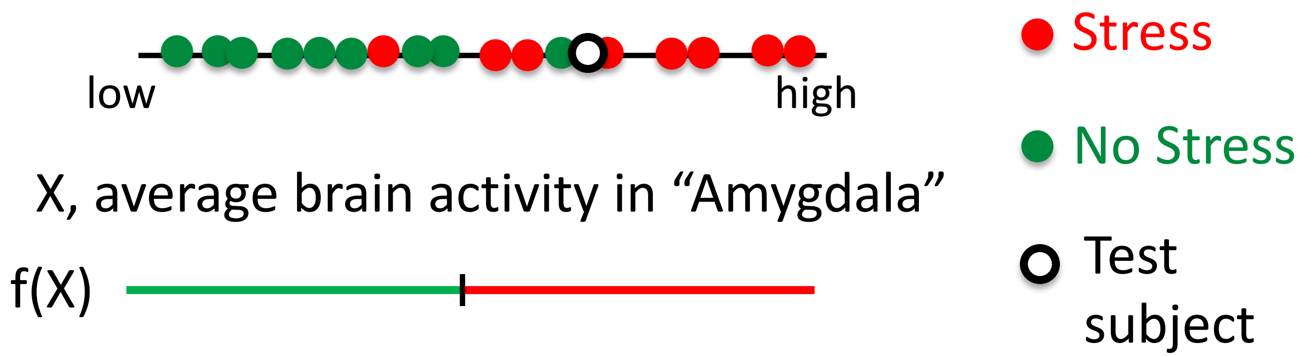
Modeling Class distribution P(Y) = Categorical(p_H,p_M,p_L)

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

Like a dice roll

$$p_H + p_M + p_L = 1$$

Modeling class conditional distribution of inputs

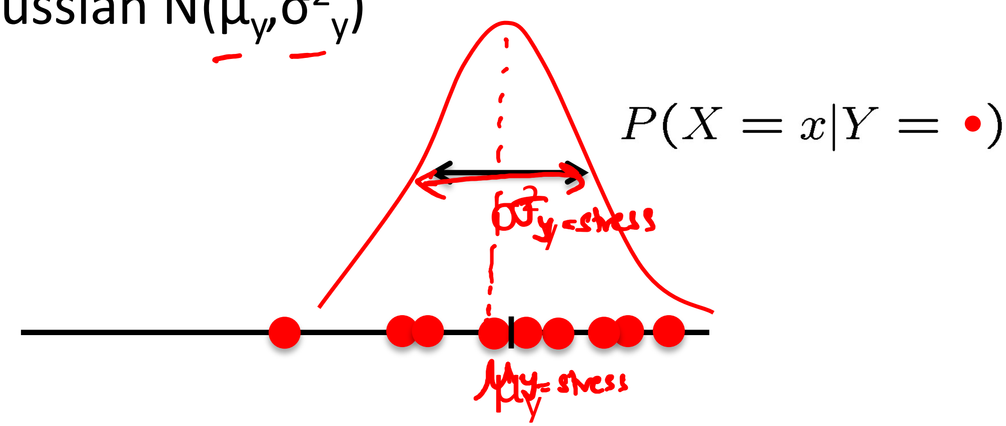


Modeling class conditional distribution of input P(X=x|Y=y)

➢ What distribution would you use?

E.g. P(X=x|Y=y) = Gaussian N(μ_y,σ^2_y)

$\mu_{Nostress}$

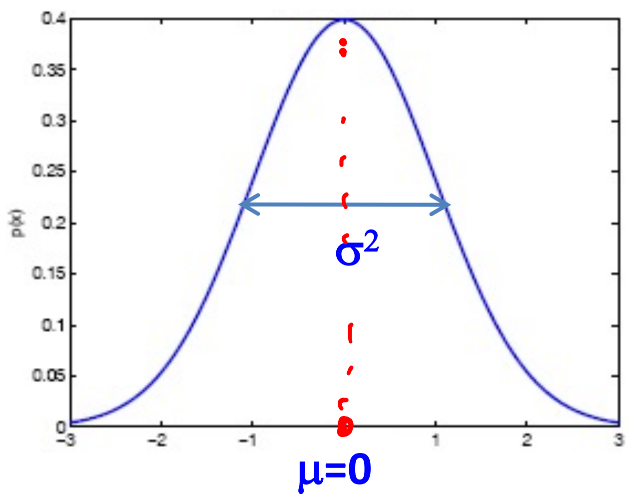


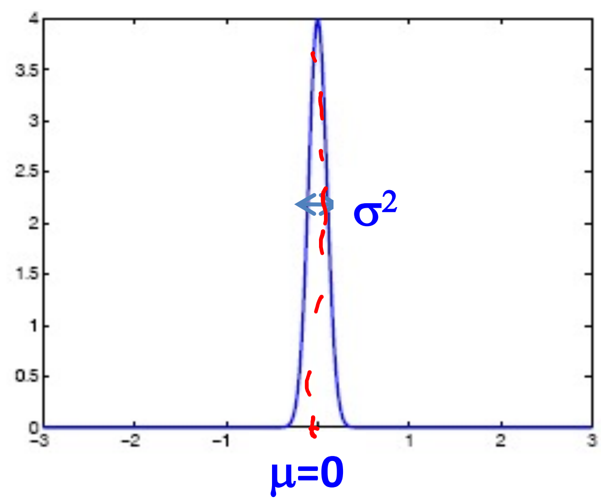
1-dim Gaussian distribution

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

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

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





1-dim Gaussian Bayes classifier

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

Class conditional
Distribution of inputs

Class distribution

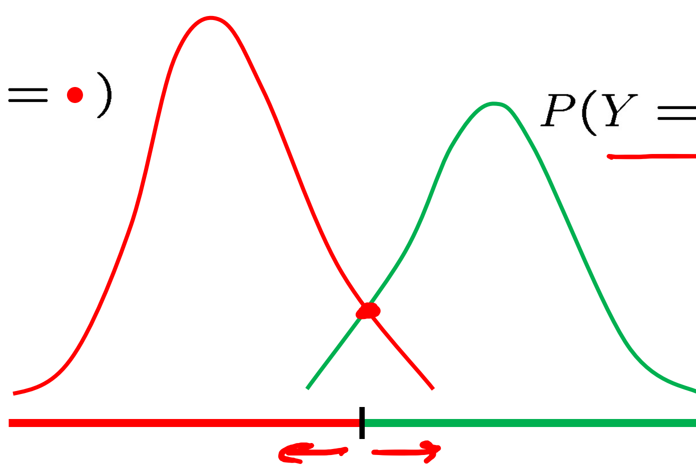
Learn parameters θ, μ_y, σ_y from data

Gaussian(μ_y, $\sigma^2{}_y$)

Bernoulli(θ)

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

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



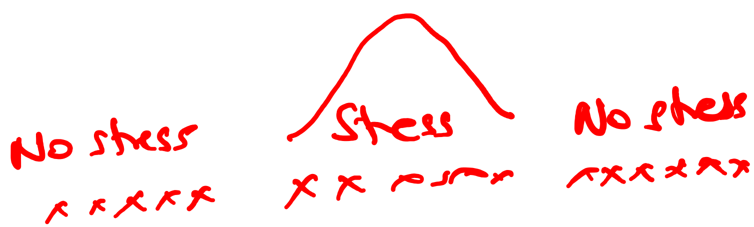
Poll

- Is the Gaussian Bayes Classifier optimal under 0/1 loss?

  A. True B. False

$P_{XY} \rightarrow$ Bayes optimal classifier $\arg\max_y P(Y|X)$

$= \arg\max_y P(X|Y)P(Y)$



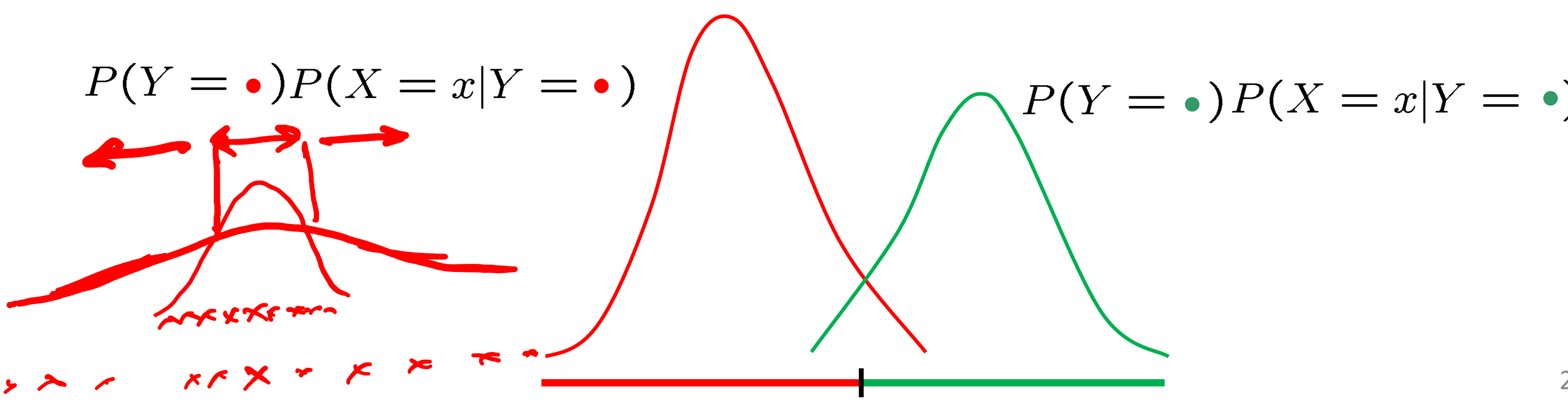
1-dim Gaussian Bayes classifier

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

Class conditional Distribution of inputs → Gaussian(μ_y, σ^2_y)

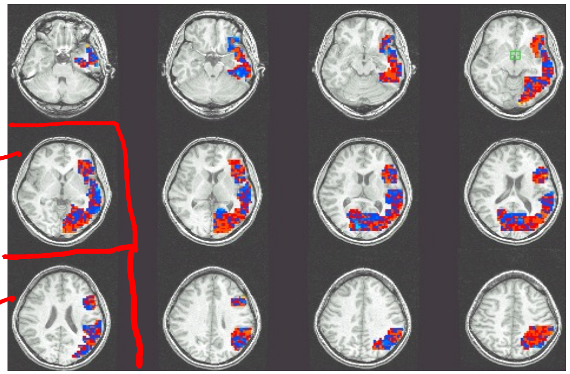
Class distribution → Bernoulli(θ)

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



d-dimensional inputs

X_2



X_1

Input feature vector, X

High Stress
Moderate Stress
Low Stress

Label, Y

$$X_1 = \begin{bmatrix} 0.05 \\ -0.06 \\ 0.1 \\ 0.3 \\ -0.1 \\ \vdots \end{bmatrix}$$

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

➢ What distribution would you use?

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

$$\mu_y = \begin{bmatrix} \\ \\ \end{bmatrix}_{d \times 1} \qquad \Sigma_y = \begin{bmatrix} \square & & \\ & \ddots & \\ & & \square \end{bmatrix}_{d \times d}$$

d-dim Gaussian distribution

$\mu = E[X]$ $\quad\quad$ $\Sigma_{ij} = E[(X_{(i)} - EX_{(i)})(X_{(j)} - EX_{(j)})]$ $\leftarrow$

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

$\Sigma_{ii} = E[(X_{(i)} - EX_{(i)})^2] \leftarrow var(X_{(i)})$

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

d-dim Gaussian distribution

X is Gaussian N(μ, Σ)

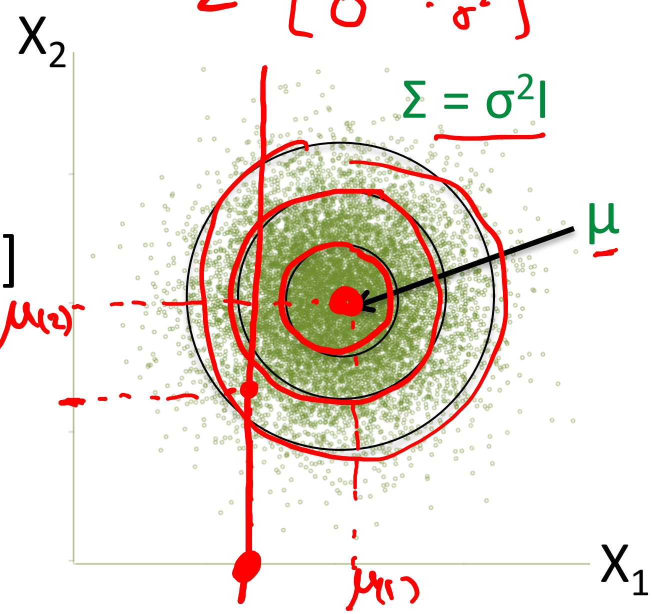
μ is d-dim vector, Σ is dxd dim matrix

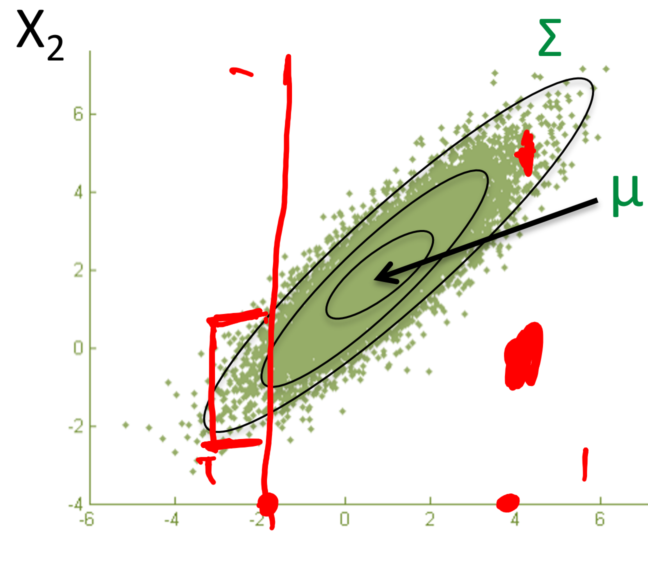
$$P(X=x|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

$$\Sigma = \begin{bmatrix} \sigma^2 & & 0 \\ 0 & \sigma^2 & \\ & \ddots & \sigma^2 \end{bmatrix}$$

d=2

$X = [X_1; X_2]$





d-dim Gaussian Bayes classifier

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

Learn parameters θ, μ_y, Σ_y from data

Class conditional

Distribution of inputs

↓

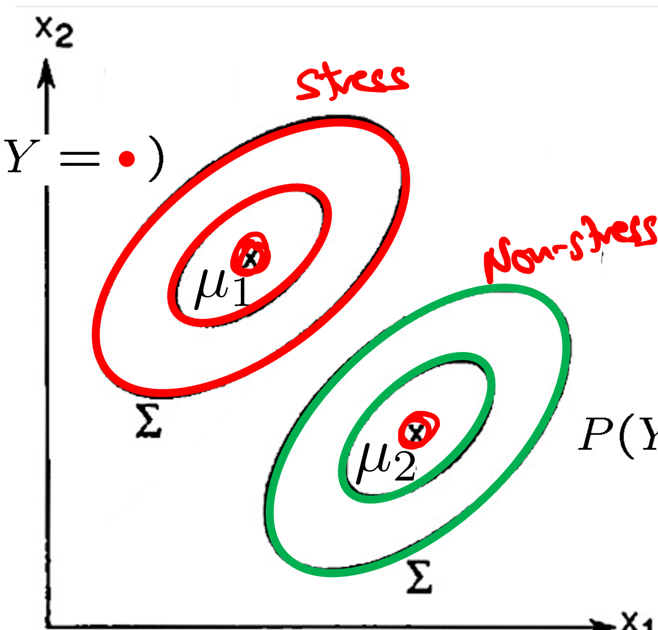
Gaussian(μ_y, Σ_y)

Class distribution

↓

Bernoulli(θ)

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



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

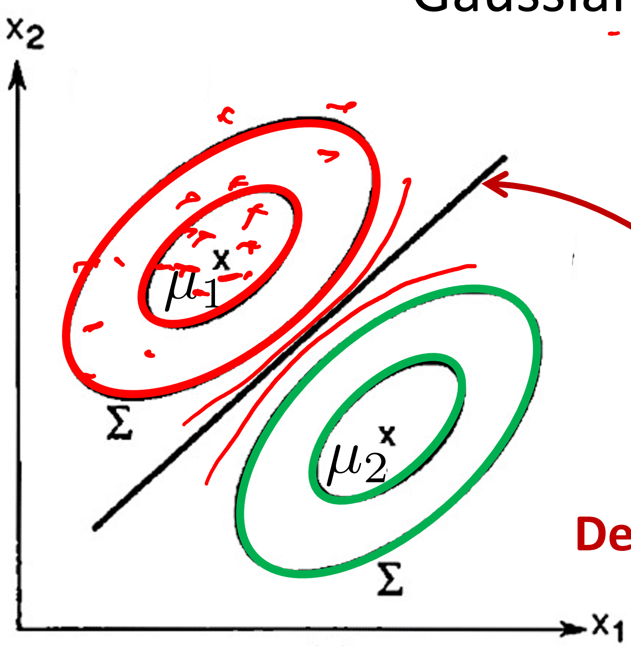
d-dim Gaussian Bayes classifier

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

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

Class distribution → Bernoulli(θ)

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



Decision Boundary