

Bayes classifier, Decision boundary

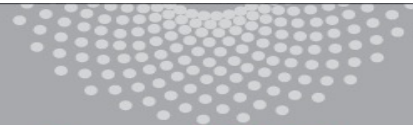
Aarti Singh

Machine Learning 10-315

Sept 1, 2021



MACHINE LEARNING DEPARTMENT



Carnegie Mellon.
School of Computer Science

In-person Expectations

- Please **do not come to class if you feel sick** or have symptoms of a potentially contagious illness.
- This holds even if you know you do not have COVID-19 but have symptoms that may be a sign of other contagious illnesses, such as a cold or flu.
- Coming to class when sick is NOT a sign of hard work. Be responsible, care for your classmates and campus community.

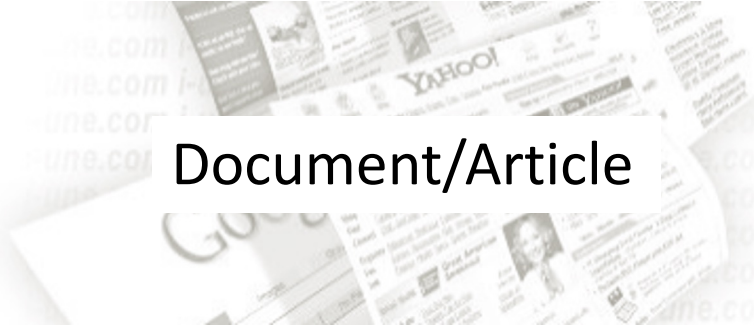
Announcements

- 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 $X \in \mathcal{X}$



Document/Article

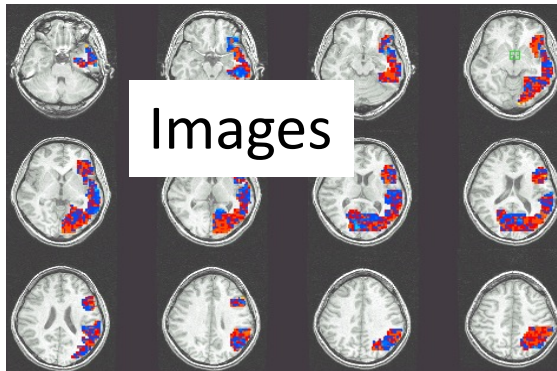
remember to wake up when class ends
=
wake ends to class remember up when

How to represent inputs mathematically?

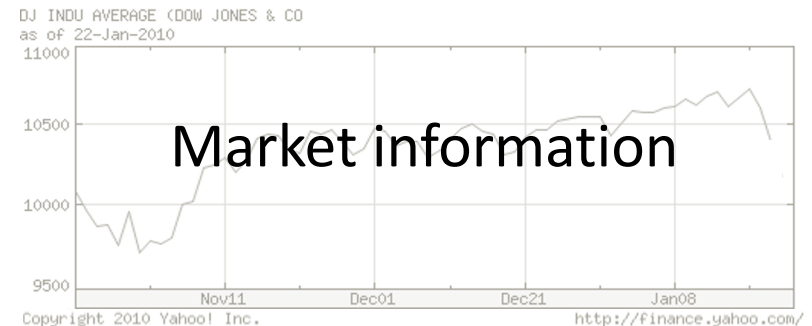
- Document vector X ➤ Ideas?
 - list of words (different length for each document)
 - frequency of words (length of each document = size of vocabulary), also known as **Bag-of-words** approach ➤ Why might this be limited?
 - list of n-grams (n-tuples of words)
- Misses out context!!

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. $P(\text{head}) = \frac{1}{2}$, $P(\text{word } x \text{ in text}) = p_x$



Probabilities in a distribution sum to 1

$$\sum_x P(X=x) = 1 \quad P(\text{tail}) = 1 - p(\text{head}), \sum_x p_x = 1$$

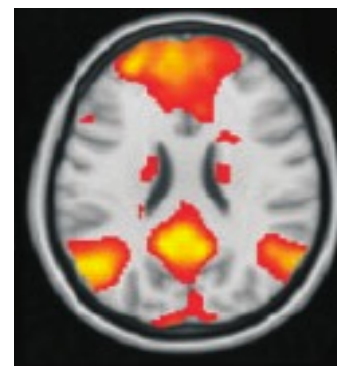
Continuous Probability density $p(x)$

e.g. $p(\text{brain activity})$

$$P(a \leq X \leq b) = \int_a^b p(x) dx$$

Probability density integrate to 1

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



Distributions in Supervised tasks

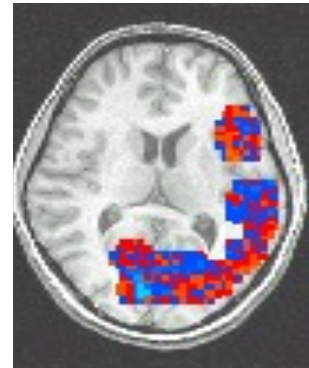
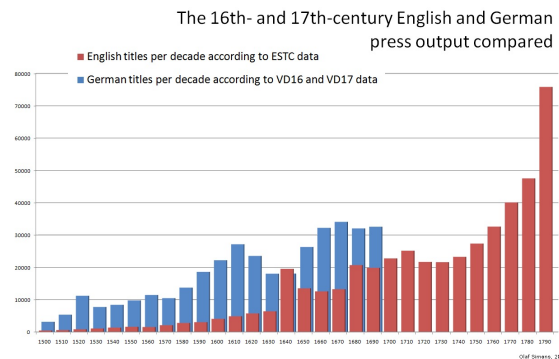
Input $X \in \mathcal{X}$

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

$P(Y = y)$ Distribution of class labels

$P(X = x \mid Y = y)$ Distribution of words in 'news' documents

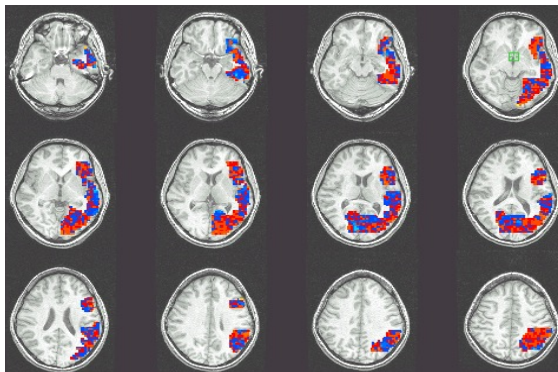
Distribution of brain activity under 'stress'



$P(Y = y \mid X = x)$ Distribution of topics given document

Classification

Goal: 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

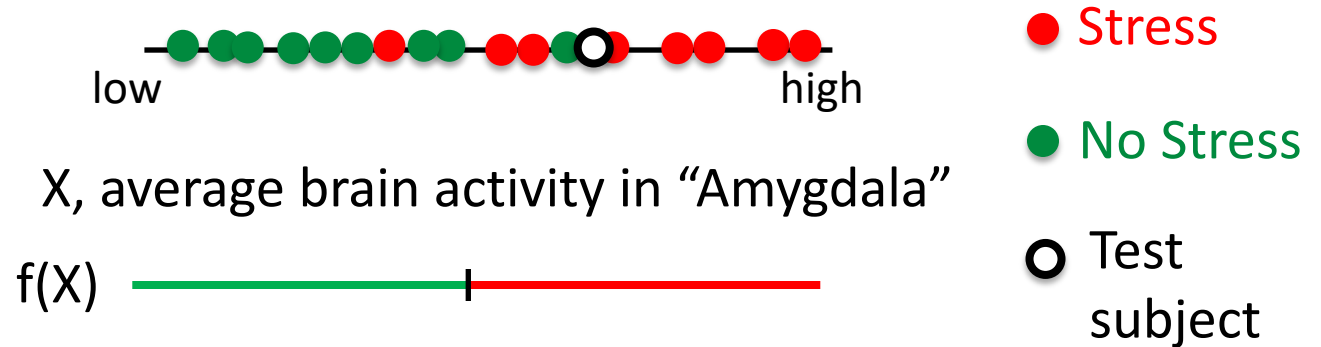
X is multi-dimensional (many features represent an input)

But let's start with a simple case:

label Y is binary (either “Stress” or “No Stress”)

X is average brain activity in the “Amygdala”

Binary Classification



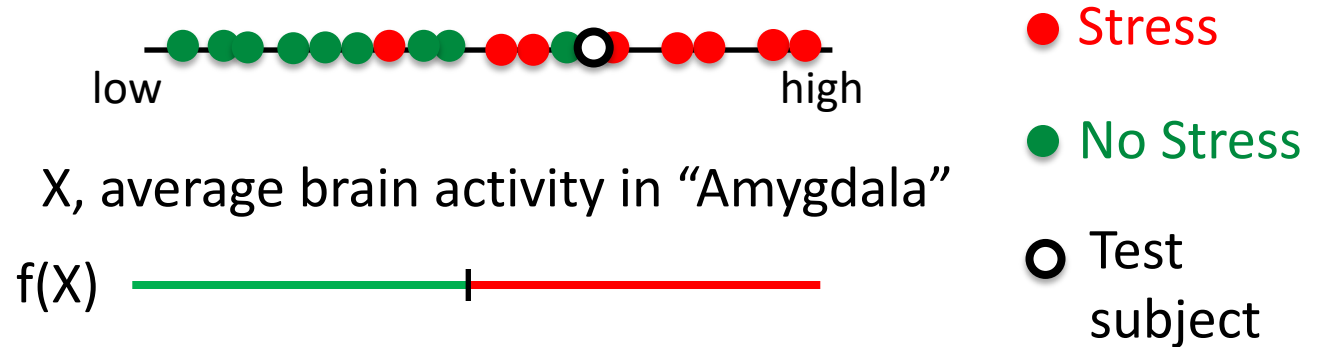
Model X and Y as random variables with joint distribution P_{XY}

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

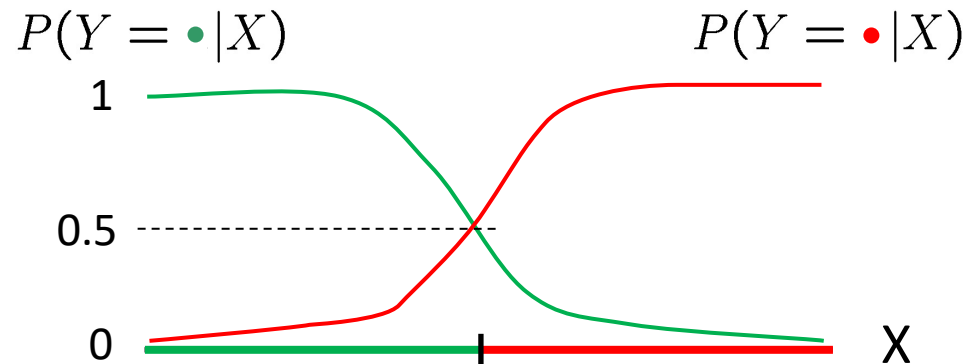
Test data $\{X, Y\} \sim \text{iid}$ sample from P_{XY}

Training and test data are independent draws from same distribution

Bayes Classifier



Model X and Y as random variables



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

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

Optimality of Bayes Classifier

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 – 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:

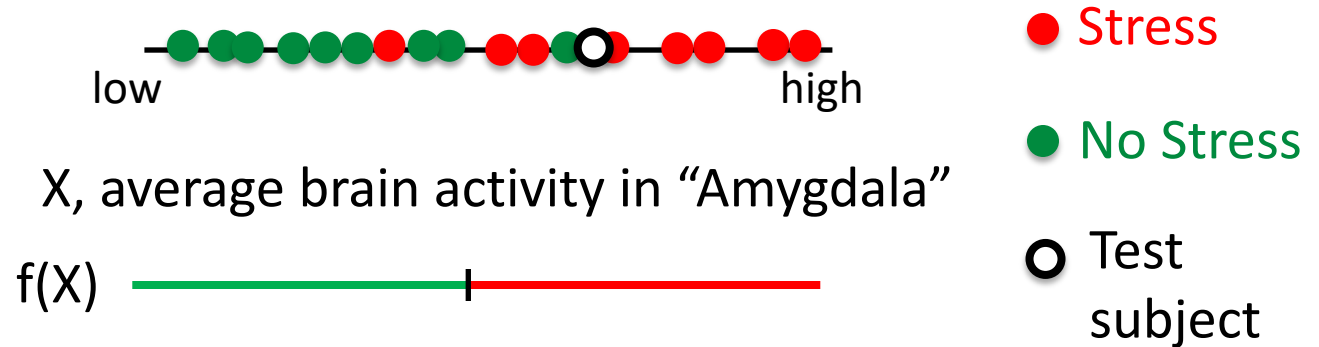
$$f(X) = \arg \max_{Y=y} P(Y = y|X = x)$$

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

Class conditional
Distribution of features

Distribution of class

Bayes Classifier



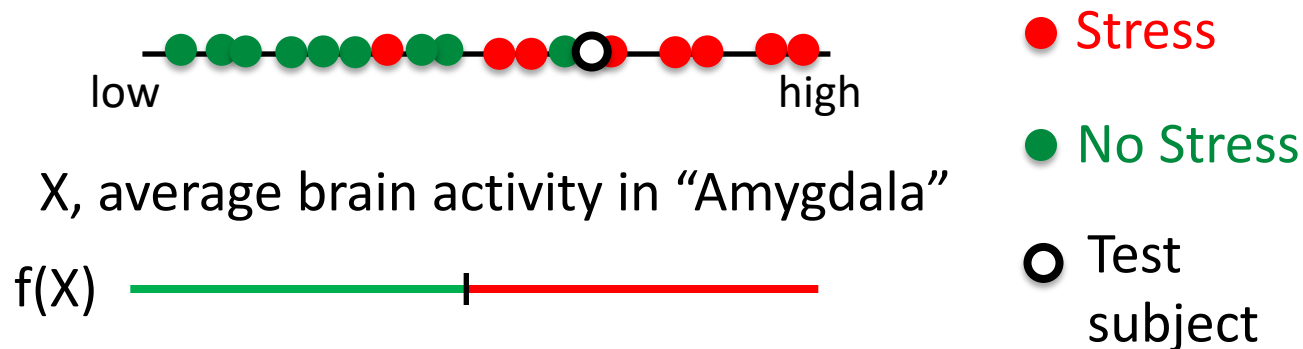
$$f(X) = \arg \max_{Y=y} \underbrace{P(X = x|Y = y)}_{\text{Class conditional Distribution of features}} \underbrace{P(Y = y)}_{\text{Class distribution}}$$

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 $P(Y=y) = \text{Bernoulli}(\theta)$

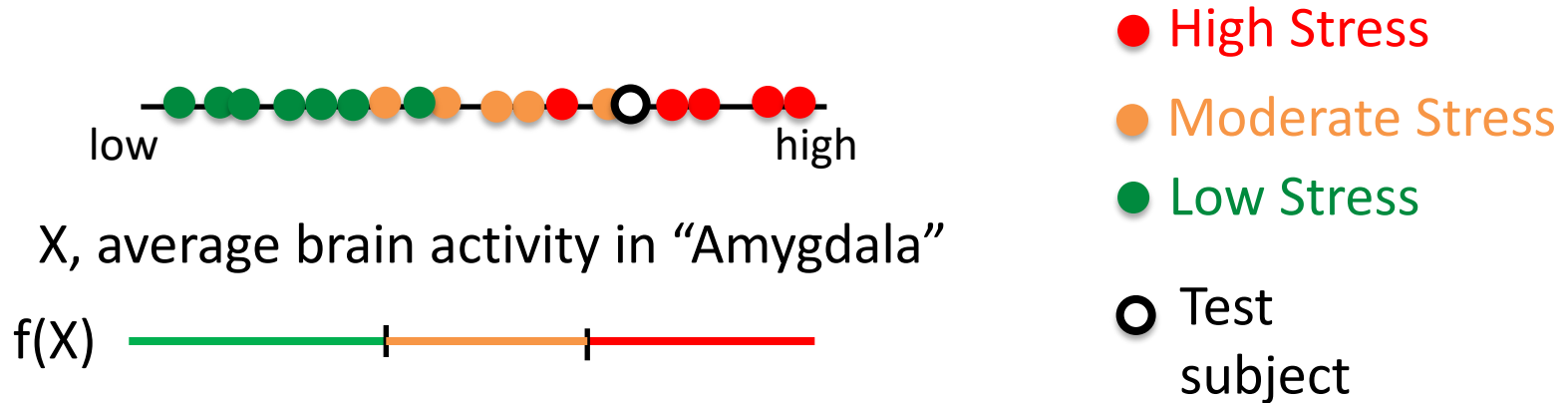
$$P(Y = \text{●}) = \theta$$

$$P(Y = \text{●}) = 1 - \theta$$

Like a coin flip



Modeling class distribution



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

Modeling Class distribution $P(Y) = \text{Multinomial}(p_H, p_M, p_L)$

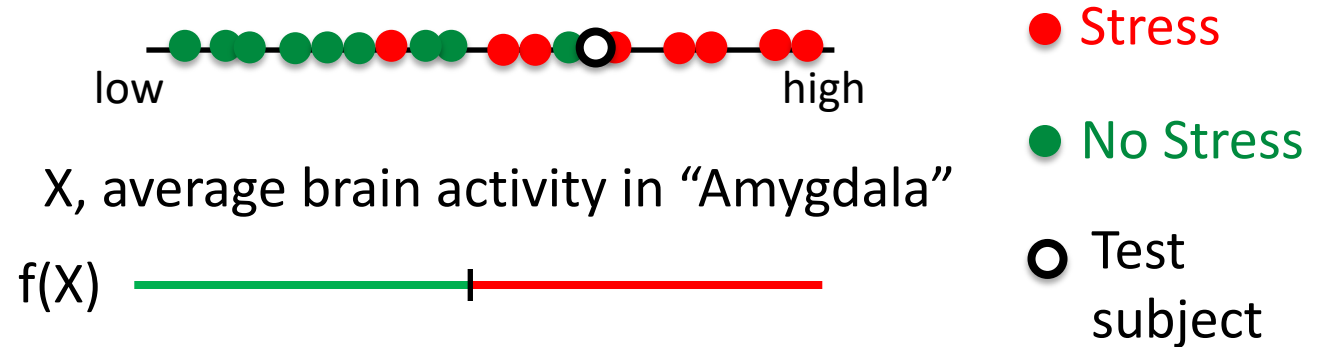
$$P(Y = \text{red}) = p_H \quad P(Y = \text{orange}) = p_M \quad P(Y = \text{green}) = p_L$$

Like a dice roll



$$p_H + p_M + p_L = 1$$

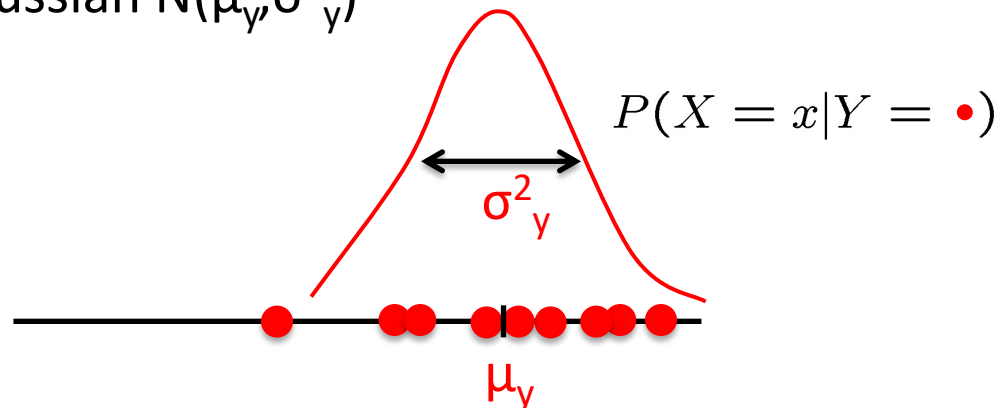
Modeling class conditional distribution of features



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

➤ What distribution would you use?

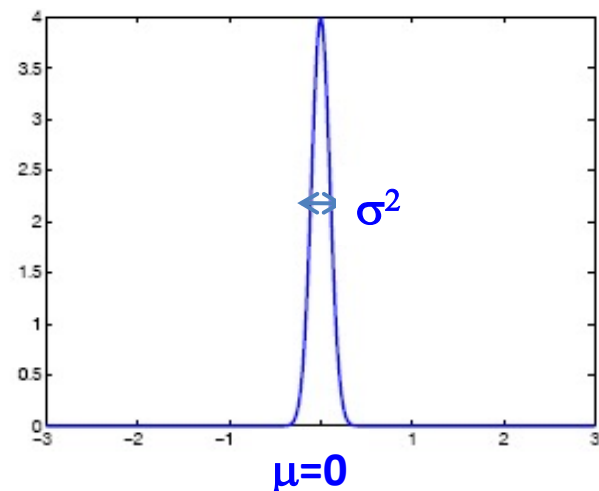
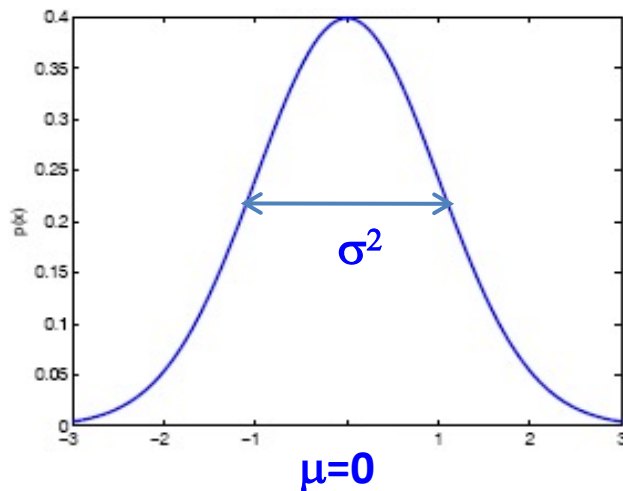
E.g. $P(X=x|Y=y) = \text{Gaussian } N(\mu_y, \sigma_y^2)$



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 $\Leftrightarrow$ Independence
 - X, Y Gaussian $\Rightarrow aX+bY$ Gaussian
 - Central limit theorem: if $X_1, \dots, X_n$ are any iid random variables with mean μ and variance $\sigma^2 < \infty$ then

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

1-dim Gaussian Bayes classifier

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

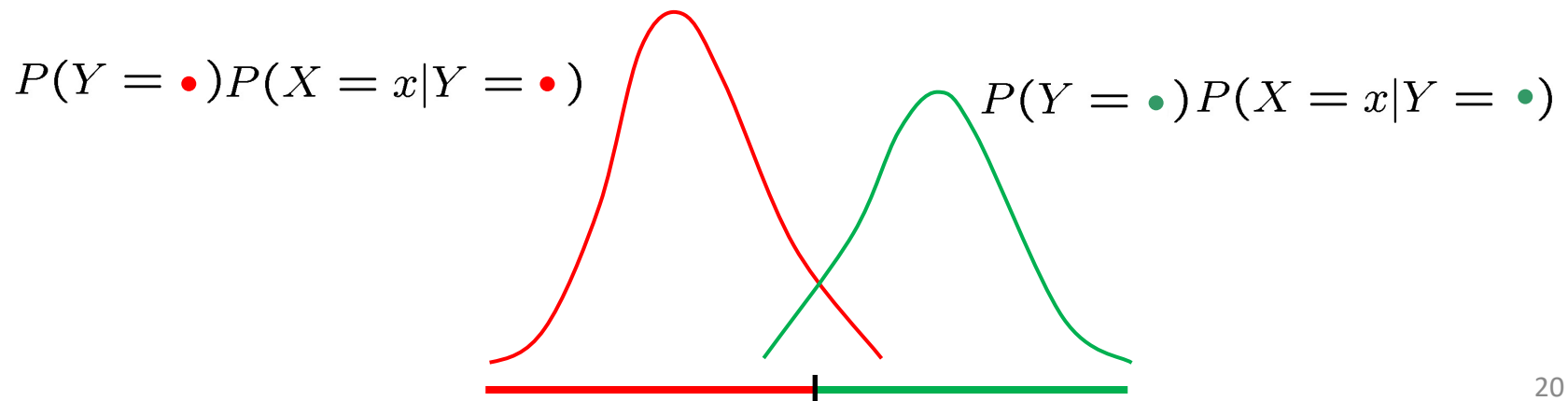
Learn parameters θ , μ_y , σ_y from data

Class conditional
Distribution of features

Class distribution

Gaussian(μ_y , σ_y^2)

Bernoulli(θ)



1-dim Gaussian Bayes classifier

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

Class conditional
Distribution of features

Class distribution

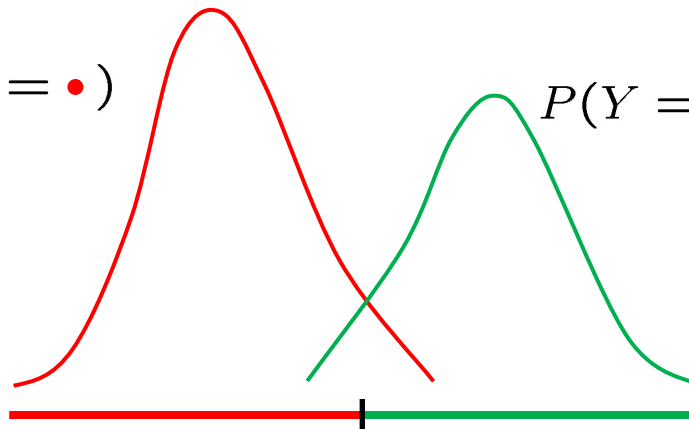
➤ What decision boundaries can we get in 1-dim?

Gaussian(μ_y, σ_y^2)

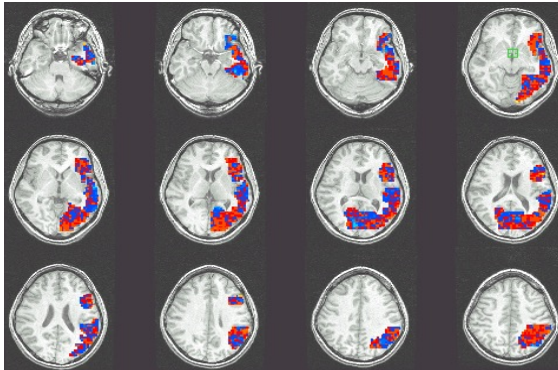
Bernoulli(θ)

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

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



d-dimensional inputs



Input feature vector, X



High Stress
Moderate Stress
Low Stress

Label, Y

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

➤ What distribution would you use?

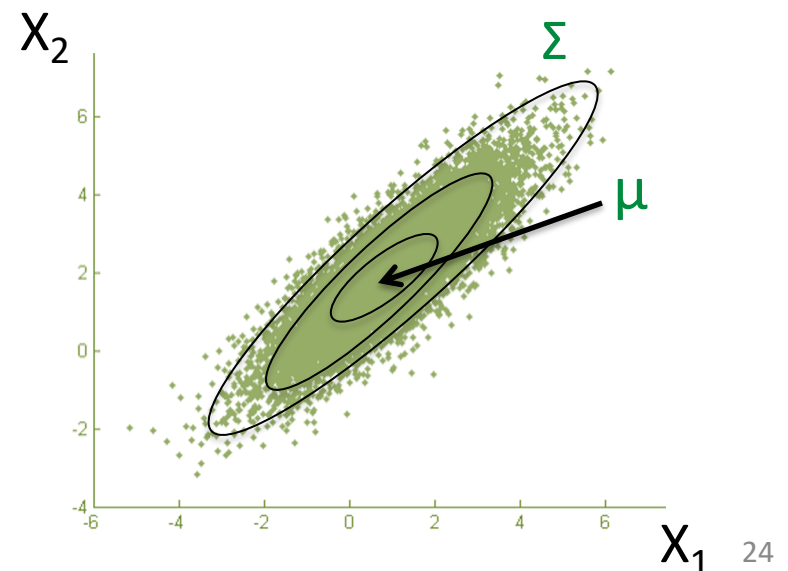
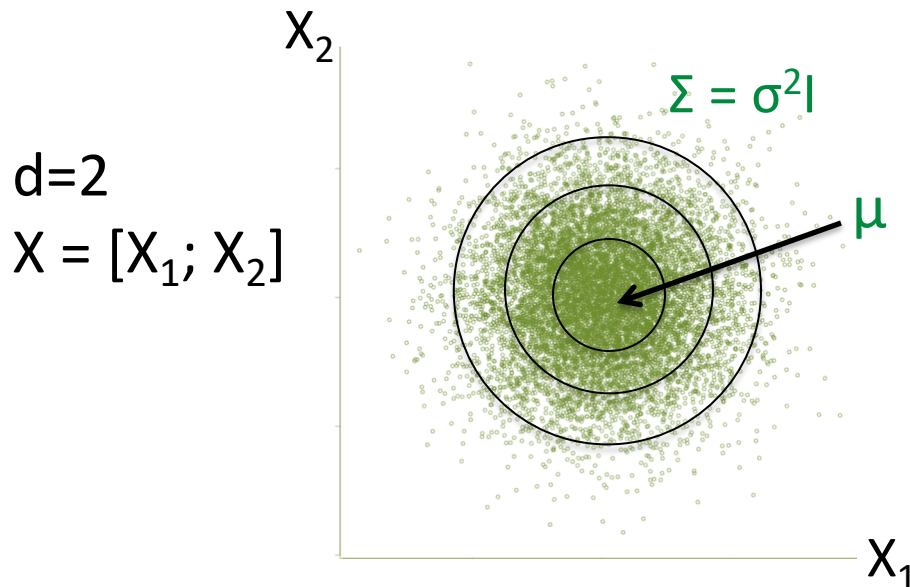
E.g. $P(X=x|Y=y) = \text{Gaussian } N(\mu_y, \Sigma_y)$

d-dim Gaussian distribution

X is Gaussian $N(\mu, \Sigma)$

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

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



d-dim Gaussian Bayes classifier

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

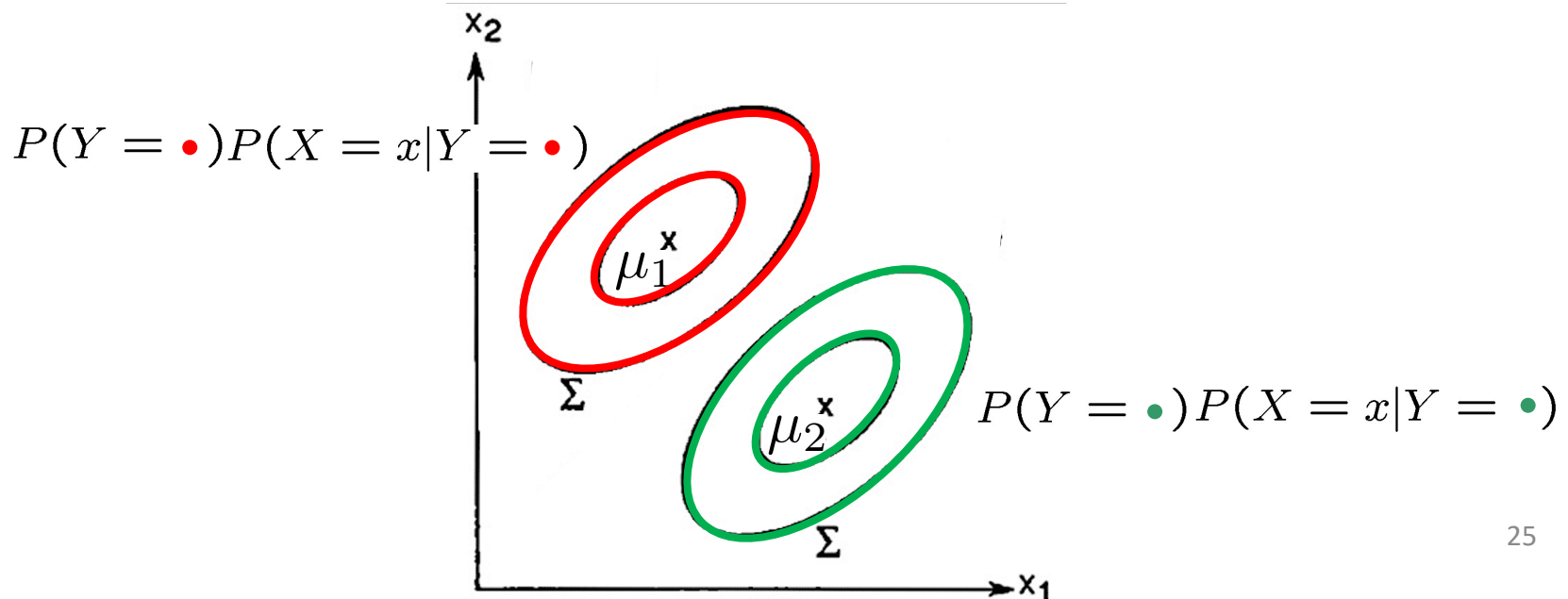
Learn parameters θ, μ_y, Σ_y from data

Class conditional
Distribution of features

Class distribution

Gaussian(μ_y, Σ_y)

Bernoulli(θ)



d-dim Gaussian Bayes classifier

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

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

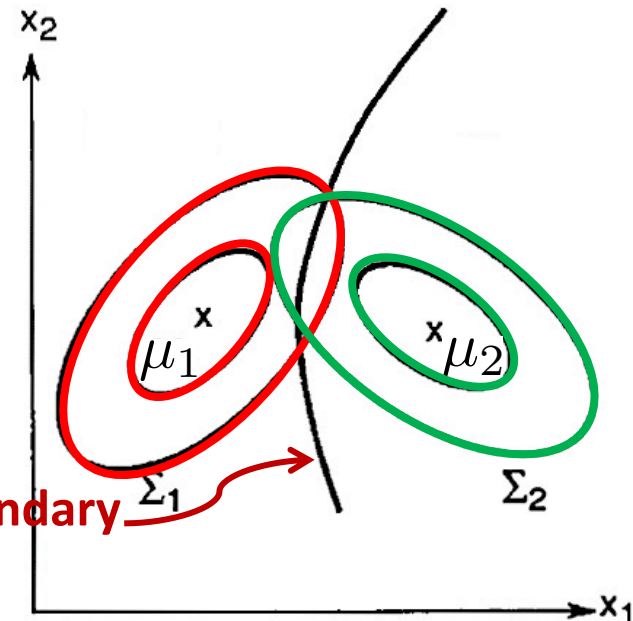
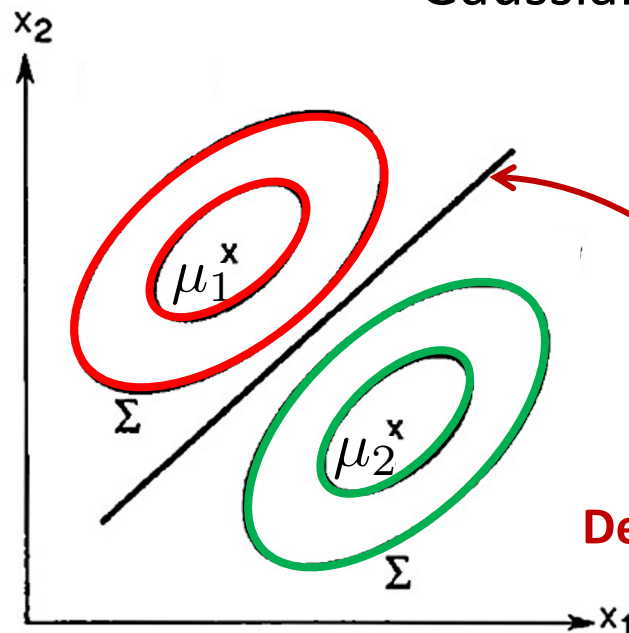
Class conditional

Class distribution

Distribution of features

Gaussian(μ_y, Σ_y)

Bernoulli(θ)



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

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

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

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.

Recap

- **Bayes classifier** – assumes P_{XY} known, optimal for 0/1 loss

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

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

Distribution of features

- **Gaussian Bayes classifier** – assumes
Class distribution is Bernoulli/Multinomial
Class conditional distribution of features is Gaussian
- **Decision boundary** – (binary classification)

Glossary of Machine Learning

- Feature/Attribute
- iid
- Bayes classifier
- Class distribution
- Class conditional distribution of features
- Decision boundary