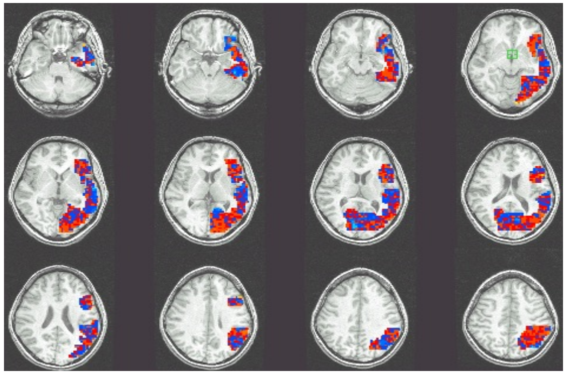
Announcements

- Recitation on Friday Jan 28 – Convexity review
- QnA1 due TODAY
- HW1 to be released TODAY

Recap – Bayes classifier



High Stress
Moderate Stress
Low Stress

(X, Y) - random variables with joint distribution P_{XY}

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

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

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

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

Class conditional Distribution of features

Class distribution

Recap – Gaussian Bayes classifier

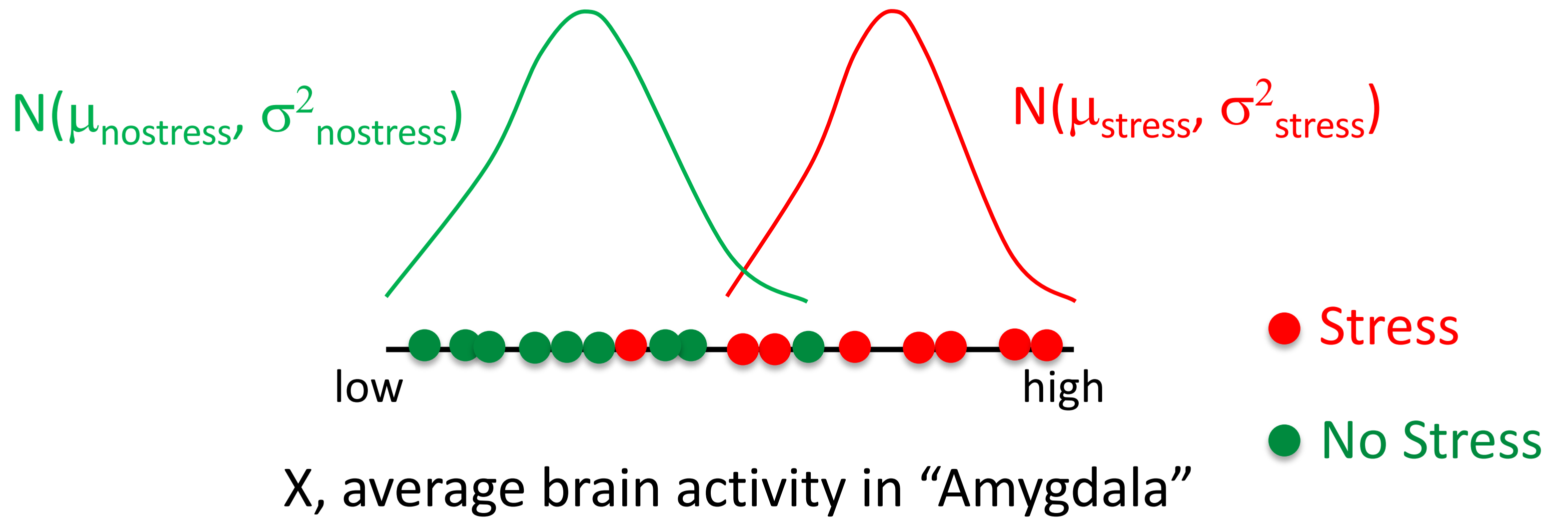
In practice P_{XY} unknown, use a distribution model to approximate

Gaussian Bayes classifier – assumes

Class distribution P(Y) is Bernoulli(θ)

[Categorical if multiple classes]

Class conditional distribution of features P(X|Y) is Gaussian



d-dim Gaussian Bayes classifier

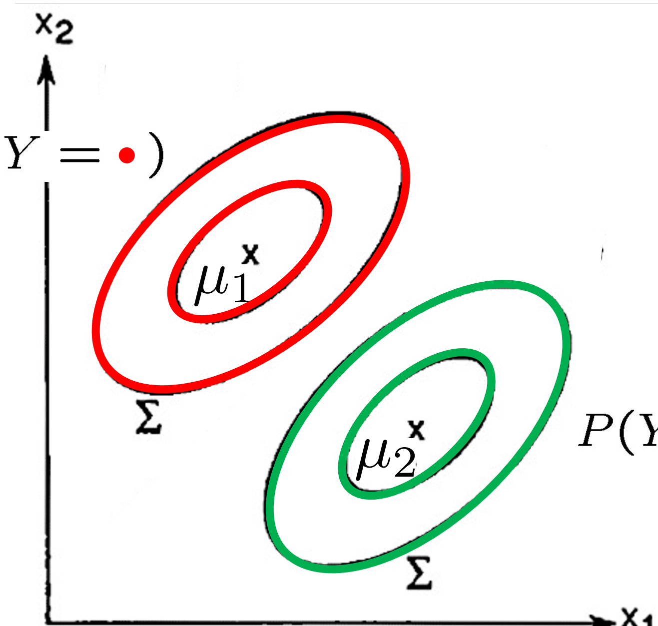
$$f(X) = \arg\max_{Y=y} P(X = x|Y = y)P(Y = y)$$

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

Class distribution → Bernoulli(θ)

Learn parameters θ, μ_y, Σ_y from data

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



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

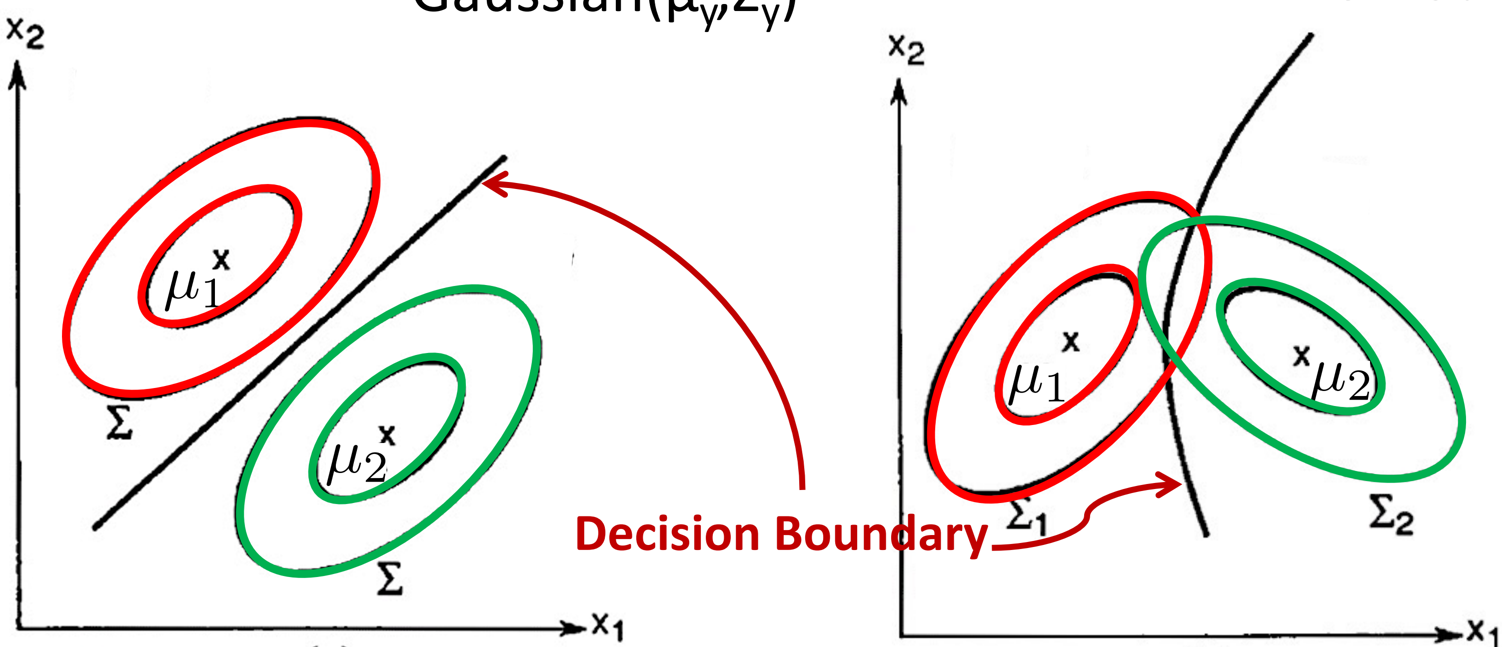
d-dim Gaussian Bayes classifier

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

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

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

Class distribution: 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

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

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

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

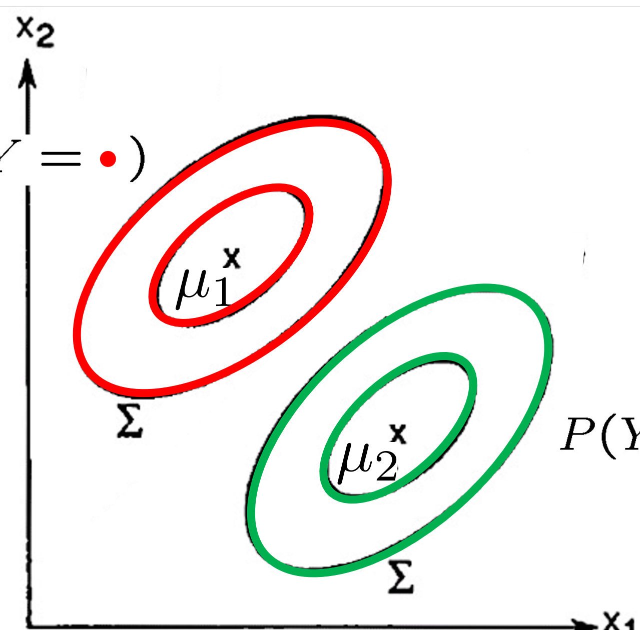
$$f(X) = \arg\max_{Y=y} P(X=x|Y=y)P(Y=y)$$

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

Class distribution → Bernoulli(θ)

Learn parameters θ, μ_y, Σ_y from data

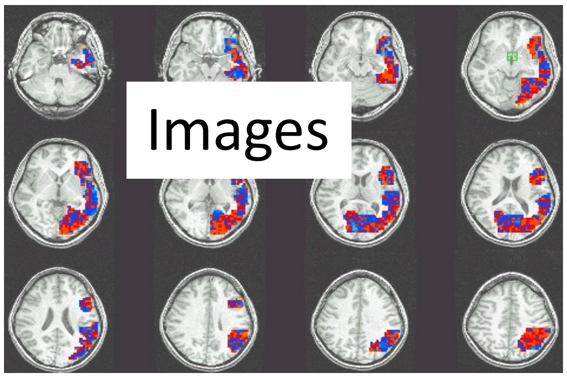
$P(Y=\bullet)P(X=x|Y=\bullet)$



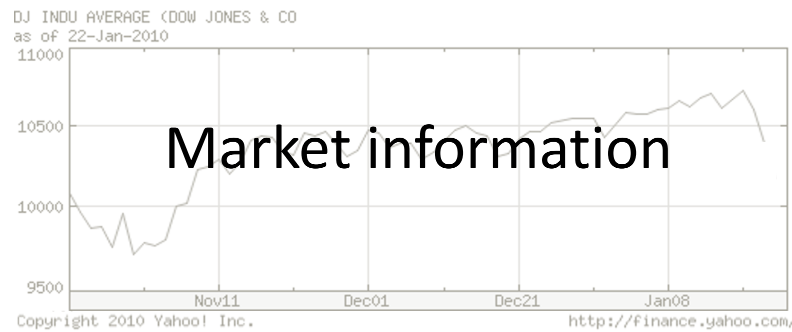
$P(Y=\bullet)P(X=x|Y=\bullet)$

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

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)

Text classification

Raw input ➡ Features ➡ Model for input features


Third Edition April 28, 2020

The Quarantine Times

Kids Connecting to their Community

Dear Future Me, Letters to remind us all what life under quarantine was like

Dear future self,
Right now we are Quarantined because of Covid-19, and we can not go anywhere! I have been reading and riding my bike so much lately! Today is Earth day, so I am going to spend lots of time outside!
Nani
3rd Grade, Edgewood Campus School

Dear 20 year old Claire,
Hi! What do you remember about Covid-19? Today I got a video that showed the 4th grade classroom. I was happy to see my class. I got to talk to my friend. She was biking across the street!
My mom and I started a newspaper called "Quarantine Times." I miss all my friends. I have a few questions for you (me). What's your job? How is it like being 20?

word1	5
word2	2
word3	10
word4	20
word5	12
word6	5
word7	8
word8	4
.	.
.	.
.	.

$$P(X=x|Y=y)$$
$$= P(\text{word1} = 5, \text{word2} = 2, \text{word3} = 10, \ldots | Y=y)$$

HW1!

Glossary of Machine Learning

- Task
- Supervised learning
  - Classification
  - Regression
- Unsupervised learning
  - Learning distribution
  - Clustering
  - Dimensionality reduction/Embedding
- Input, X
- Label, Y
- Prediction, f(X)
- Experience = Training data
- Test data
- Overfitting
- Generalization
- Performance measure/loss – 0/1, squared
- iid
- Class conditional distribution of inputs
- Bayes rule
- Bayes Optimal classifier
- Decision boundary
- Feature/Attribute

Maximum Likelihood Estimation (MLE)

Aarti Singh

Machine Learning 10-315
Jan 26, 2022

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 = 

- Parameter θ : 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)

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

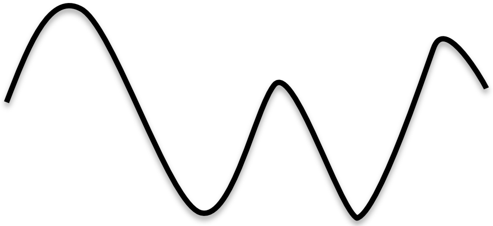
MLE of probability of head:

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

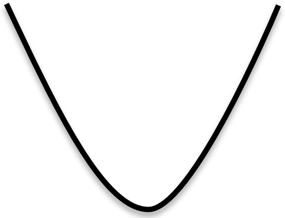
"Frequency of heads"

Short detour - Optimization

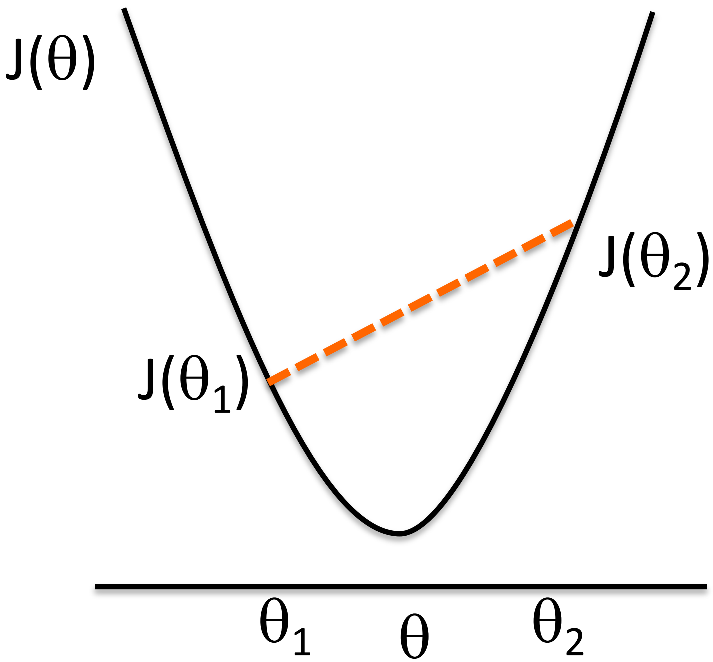
- Optimization objective $J(\theta)$
- Minimum value $J^* = \min_\theta J(\theta)$
- Minima (points at which minimum value is achieved) may not be unique



- If function is strictly convex, then minimum is unique

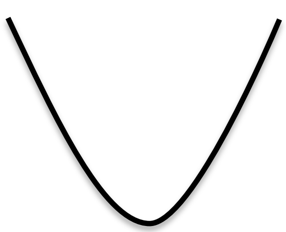


Convex functions

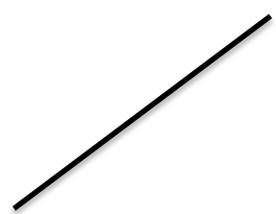


A function $J(\theta)$ is called **convex** if the line joining two points $J(\theta_1), J(\theta_2)$ on the function does not go below the function on the interval $[\theta_1, \theta_2]$

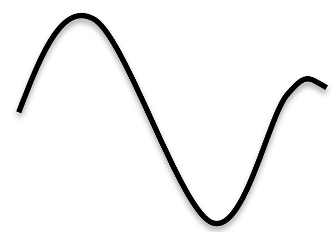
(Strictly) Convex functions have a unique minimum!



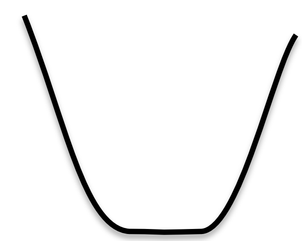
Convex



Both Concave & Convex



Neither



Convex but not strictly convex

Optimizing convex (concave) functions

- Derivative of a function

- Derivative is zero at minimum of a convex function

- Second derivative is positive at minimum of a convex function

Bernoulli MLE Derivation

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

Categorical distribution

Data, D = rolls of a dice

- $P(1) = p_1, \; P(2) = p_2, \ldots, P(6) = p_6 \quad p_1 + \ldots + p_6 = 1$
- Rolls are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Categorical(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:

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

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

α_y ← Rolls that turn up y

$\sum_y \alpha_y$ ← Total number of rolls

"Frequency of roll y"

How to learn parameters from data? MLE

(Continuous case)

d-dim Gaussian distribution

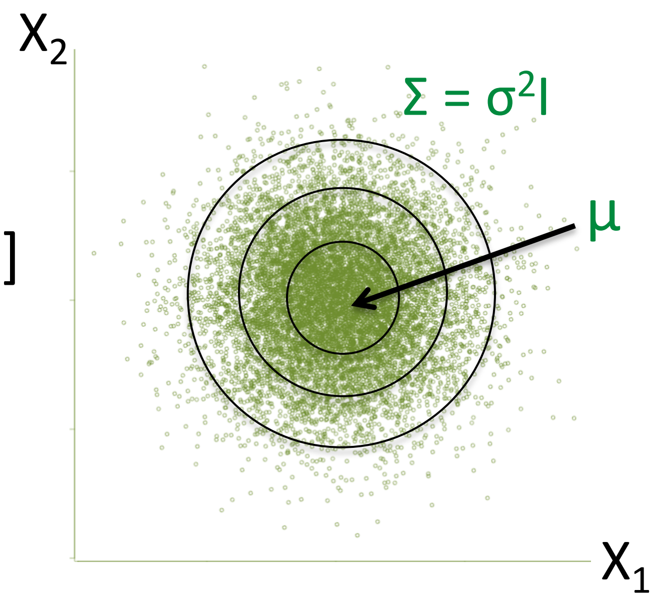
X is Gaussian N(μ, Σ)

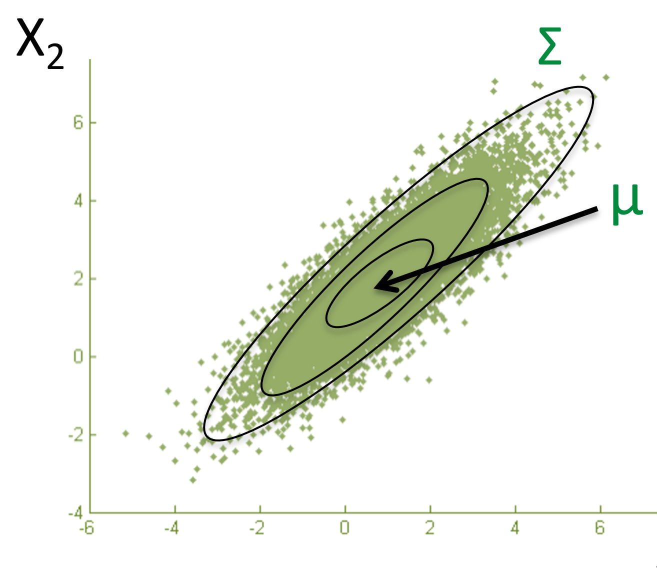
μ is d-dim vector, Σ is dxd dim matrix

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

d=2

$X = [X_1; X_2]$

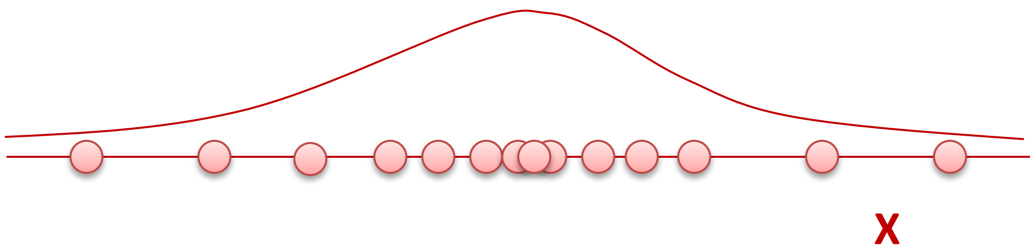




How to learn parameters from data? MLE

(Continuous case)

Gaussian distribution

Data, D = 

- Parameters: μ – mean, σ^2 - variance

- Data 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 θ= (μ,σ²) that maximizes the probability of observed data

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

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

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

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

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

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

MLE for Gaussian mean

- Poll

P(D|θ) = $$\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}$$

A. $$\max_{\mu} \sum_{i=1}^{n} (X_i - \mu)^2$$

B. $$\min_{\mu} \sum_{i=1}^{n} (X_i - \mu)^2$$

C. $$\max_{\mu} \mu^2 - 2\mu \sum_{i=1}^{n} X_i$$

D. $$\max_{\mu} n\mu^2 - 2\mu \sum_{i=1}^{n} X_i$$

MLE for Gaussian mean and variance

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

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

Self exercise:

Derive MLE of variance?

d-dimensional versions?

MLE for uniform or exponential distribution?

More coming up in HW1