Announcements

- Anonymous feedback form
- Recitation on Friday Sept 10 – Convexity review
- QnA1 due TODAY
- HW1 to be released TODAY

Recap

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

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

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

Class conditional
Distribution of features

Class distribution

- **Gaussian Bayes classifier** – assumes
  - Class distribution is Bernoulli/Multinomial
  - Class conditional distribution of features is Gaussian

- **Decision boundary** – (binary classification)

d-dim Gaussian Bayes classifier

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

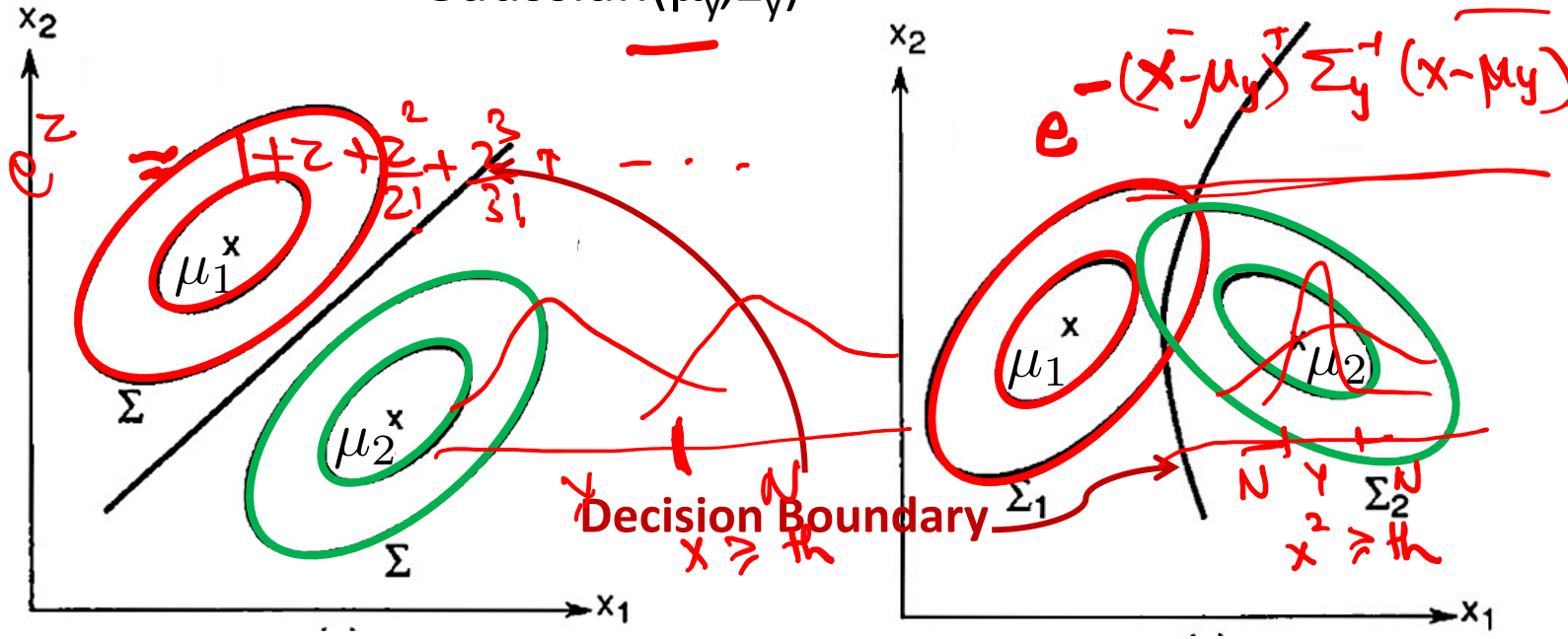
- What decision boundaries can we get in d-dim?

Class conditional Distribution of features

Class distribution

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.

Glossary of Machine Learning

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

Maximum Likelihood Estimation (MLE)

Aarti Singh

Machine Learning 10-315
Sept 8, 2021

How to learn parameters from data?
MLE

(Discrete case)

Learning parameters in distributions

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

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

Choose θ that maximizes the probability of observed data aka Likelihood

Maximum Likelihood Estimation (MLE)

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

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

$D = X_1 \ldots X_n$ (H, T) $X_i = \begin{cases} 1 & H \\ 0 & T \end{cases}$

$= P(X_1 \ldots X_n ; \theta)$

$= \prod_{i=1}^{n} P(X_i ; \theta)$

$= \prod_{i=1}^{n} \theta^{X_i} (1-\theta)^{1-X_i} = J(\theta)$

$\max_{\theta}$

MLE of probability of head:

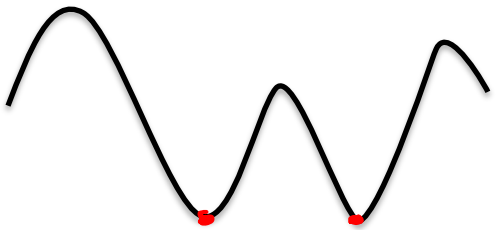
$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} = 3/5$$

(α_H: # heads)

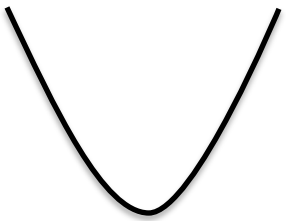
"Frequency of heads"

Short detour - Optimization

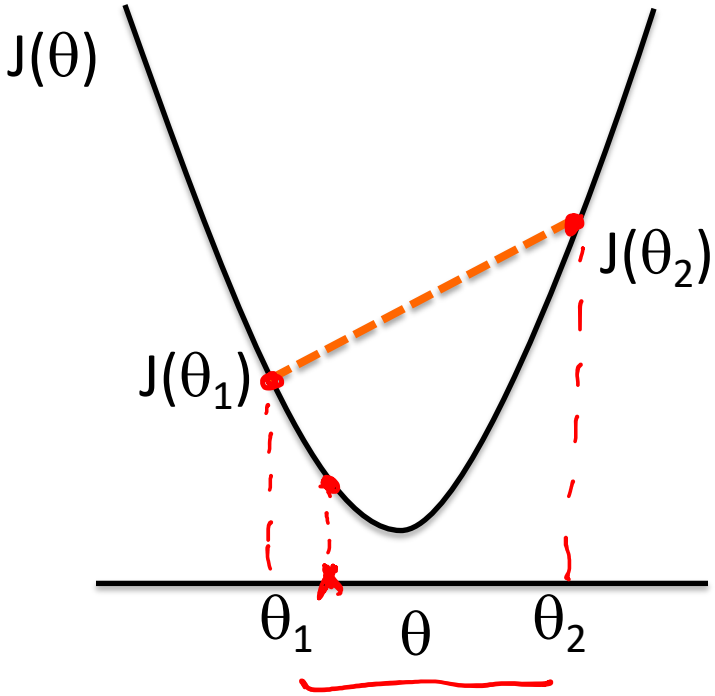
- Optimization objective $J(\theta)$
- Minimum value $J^* = \min_\theta J(\theta)$
- Minima (points at which minimum value is achieved) may not be unique



- If function is strictly convex, then minimum is unique

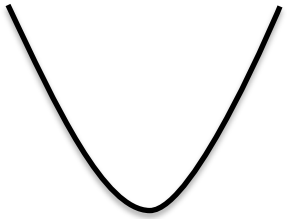


Convex functions

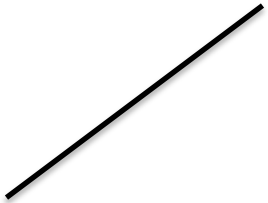


A function J(θ) is called **convex** if the line joining two points J(θ1),J(θ2) on the function does not go below the function on the interval [θ1, θ2]

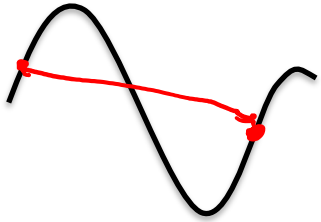
(Strictly) Convex functions have a unique minimum!



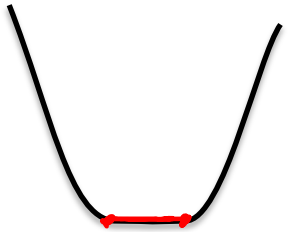
Convex



Both Concave & Convex



Neither

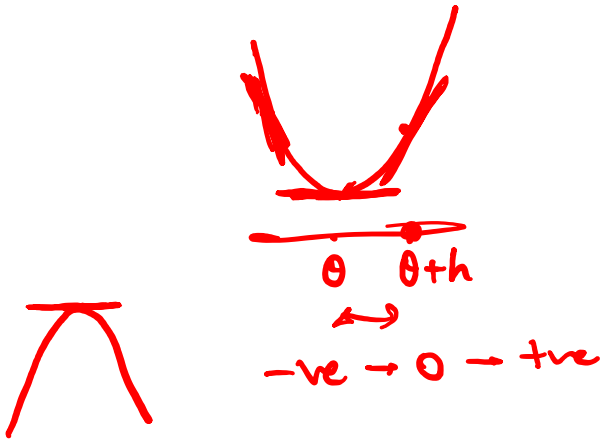


Convex but not strictly convex

Optimizing convex (concave) functions

- Derivative of a function

$$\frac{\partial J(\theta)}{\partial(\theta)} = \lim_{h \to 0} \frac{J(\theta+h) - J(\theta)}{h}$$



  - 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?

Bernoulli MLE Derivation

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

$$= \arg\max_{\theta} \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}$$

$$= \arg\max_{\theta} \underbrace{\sum_{i=1}^{n} x_i \log\theta + (1-x_i)\log(1-\theta)}_{J(\theta)}$$

$$\rightarrow \frac{\partial J(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{x_i}{\theta} + \frac{(1-x_i)}{1-\theta} = 0$$

$$x_i = \begin{cases} 1 & H \\ 0 & T \end{cases}$$

$$\frac{\sum_{i=1}^{n} x_i}{\theta} = \frac{\sum_{i=1}^{n}(1-x_i)}{1-\theta}$$

$$\frac{\alpha_H}{\theta} = \frac{\alpha_T}{1-\theta} \Rightarrow \alpha_H(1-\theta) = \alpha_T\theta$$

$$\alpha_H = (\alpha_H + \alpha_T)\theta$$

$$\theta = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

Multinomial distribution

Data, D = rolls of a dice

- $P(1) = p_1$, $P(2) = p_2$, …, $P(6) = p_6$ $p_1+….+p_6 = 1$
- 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

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

MLE of probability of rolls:

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

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

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

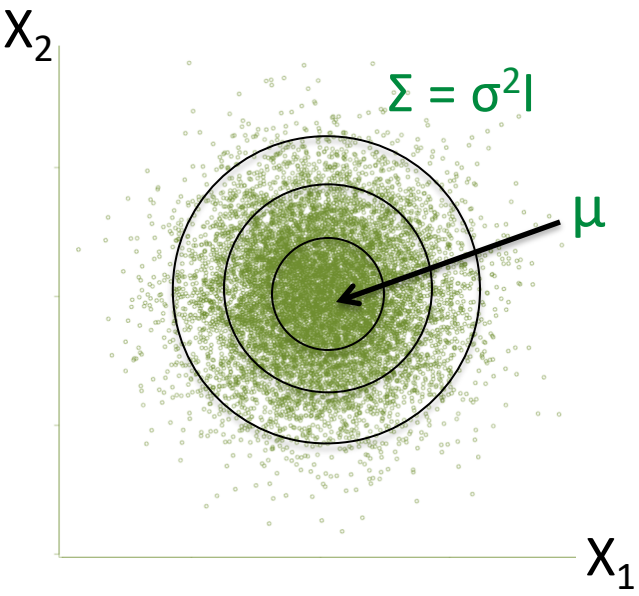
X is Gaussian N(μ, Σ)

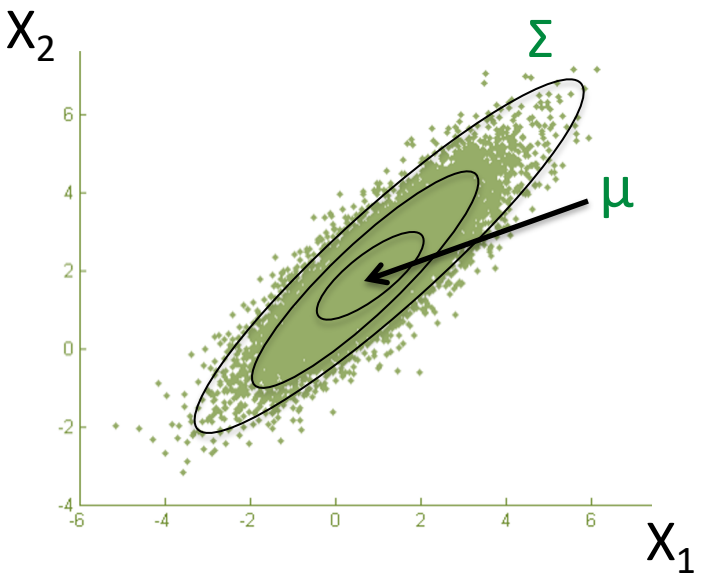
μ is d-dim vector, Σ is dxd dim matrix

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

d=2

X = [X_1; X_2]



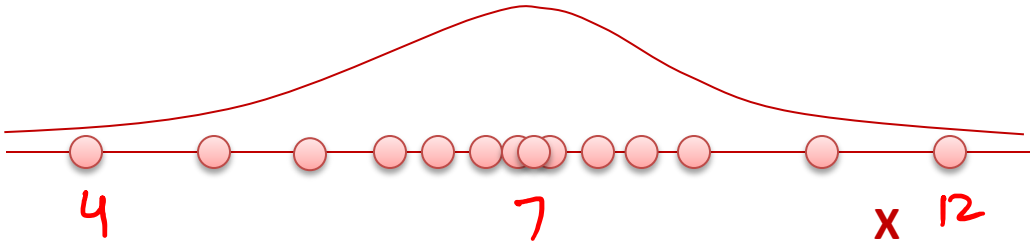


How to learn parameters from data? MLE

(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

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

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

Independent draws

$= \; \hookrightarrow \mu, \sigma^2 \qquad \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$

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

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

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

$e^a e^b = e^{a+b}$

MLE for Gaussian mean

- Poll

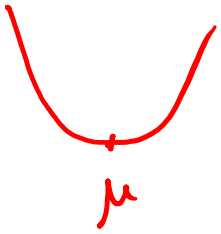
$$P(D|\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}$$

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



$$-\sum_{i=1}^{n} 2(x_i - \mu) = 0$$

$$n\mu = \sum_{i=1}^{n} x_i$$

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

MLE for Gaussian mean and variance

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

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

Self exercise:

Derive MLE of variance?

d-dimensional versions?

MLE for uniform or exponential distribution?

More coming up in HW1

Unbiased estimates

- Bias of an estimate $E[\hat{\theta}] - \theta$
- Unbiased estimate if $E[\hat{\theta}] = \theta$

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E[\hat{\mu}_{MLE}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i]$$

$$= \frac{1}{n}\, n\, \mu$$

$$= \mu$$

  - Is the MLE of mean unbiased?
  - Is the MLE of variance unbiased?

$$E[\hat{\sigma}^2_{MLE}] = \sigma^2 \frac{n-1}{n}$$

  - How can you make it unbiased?