Today
- Admin
- Gaussians
- GDA
- Linear regression
- Pick up assigment

1. Assignment 1 DUE NOW! (or Monday for 25% off)
2. No tutorial this week (Ass 2 out this weekend) (Tutorials are optional, not excuse to leave classes early)
3. Piazza (closing soon, find friends)
4. End of class → pick up assignment to mark

No

DUE

25%

his week

eekend)

2 classes

find friends)

assignment to mark

NOW!

off)

early)

Notation

I used:

D: data $\{(x_i, y_i)\}_{i=1}^{N}$

D_i: data sample (x_i, y_i)

D: dimension

@: parameter of Bernoulli

@: hypothesis for MLE

?Should we use:

DS

DS_i

P

$h \in H$

X, x_i: design matrix, features of example i

A, a_i

Y, y_i: target vector, label of example i

B, b_i

Naive

Issue

- div

- instr

pres

Naive Bayes

Issues:

- div by 0
- instructor badly presenting MLE (sorry)

Advantages

- simple
- fast (single pass) training
- parrallel
- small # params
- not much data needed

Disadvantages

- strong mutual independence assumption.
  - → limited modeling power

Course project idea:

- explore different ways to relax NB assumption (TAN-Bayes, Bayesian Network Classifiers

Today:

- What if we have $X_i \in \mathbb{R}^D$?

an | Network Classifiers)

Gaussian Distribution

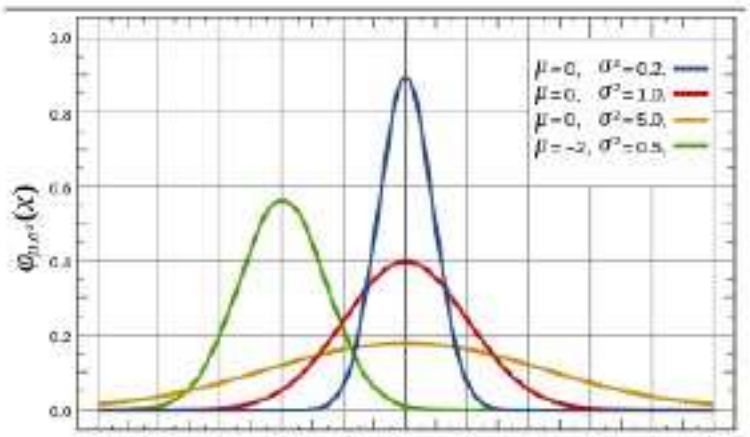
If x is distributed according to a Gaussian/normal distribution with mean μ and variance σ, we write this as

$$x \sim \mathcal{N}(\mu, \sigma^2),$$

and the probability density function (PDF) is

$$p(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

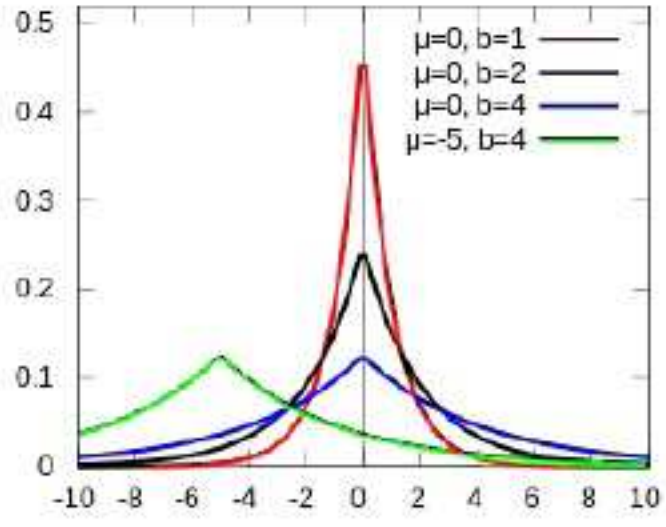
Here are some example PDFs from Wikipedia,



An alternative distribution is the Laplace distribution where we have

$$p(x|\mu, \sigma) = \frac{1}{2\sigma} \exp\left(\frac{|x-\mu|}{\sigma}\right),$$

It's Wikipedia graphic looks like this:



The Laplace distribution has 'heavier-tails' than the Gaussian distribution, meaning that it assigns a higher probability to events that are far away from μ (while the Gaussian assumes everything is close to μ). A distribution with even heavier tails is the student's t distribution.

Motivation

- Central limit theorem. → MLE
- data is Gaussian
- Analytic properties
  - sampling
  - symmetry
- $(-\infty, \infty)$

MLE

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

$p(x^{(i)}|y_i)$

$$\bar{x} \sim N(\mu, \Sigma)$$

$$\int_{-\infty}^{\infty} p(x \mid \mu, \Sigma) = 1$$

MLE

$$p(\bar{x} \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma\right.$$

$$p(\bar{x}$$

$$\mu \in \mathbb{R}^d$$

$$\Sigma \in S_{++}^d$$

$$\Sigma > 0$$

$$\hat{\mu} =$$

$$\hat{\Sigma} =$$

$$(x_i - \hat{\mu})^2$$

$$(x^i \mid y_i)$$

$$\int_{-\infty}^{\infty} p(x|m,\Sigma) = 1$$

$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

$\mu \in \mathbb{R}^d$

$\Sigma \in \mathbb{S}^d_{++}$

$\Sigma \succ 0$

$p(\bar{x}_i | y_i)$

$\bar{x}_i | y_i \sim N(m, \Sigma)$

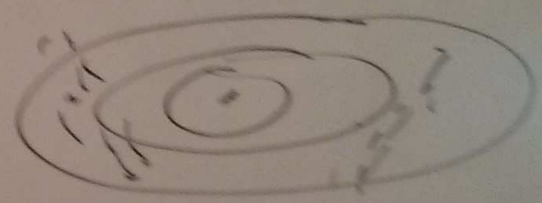
$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \bar{x}_i$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} (\bar{x}_i - m)(\bar{x}_i - m)^T \in \mathbb{S}^d_+$$

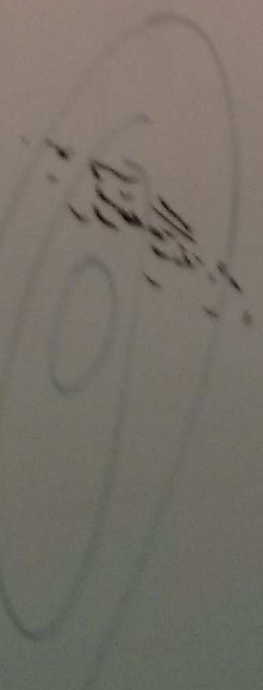
Problems:

1. $\hat{\Sigma}$ might not be valid

2. unimodal



3. not "robust"



Today

- Admin
- Gaussians
- GDA
- Linear regression
- Pick up assigment

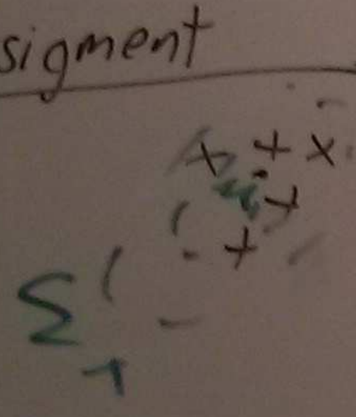
Gaussian Discriminant Analysis

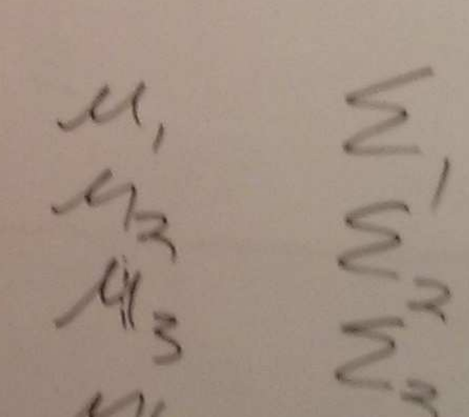
$p(\bar{x}_i, y_i) = p(\bar{x}_i | y_i) p(y_i)$

$\bar{x}_i \in \mathbb{R}^d$

$y_i \in \{1, 2, 3, 4\}$

$\bar{x}_i | y_i = c \sim N(\mu_c$





Σ_1 Σ_2 Σ_3 Σ_4

μ_1 μ_2 μ_3 μ_4

$$p(y_i | \bar{x}_i) = \frac{p(\bar{x}_i | y_i) p(y_i)}{p(\bar{x}_i)} \propto p(\bar{x}_i | y_i) p(y_i)$$

$$p(\bar{x}_i) = \sum_{c=1}^{C} p(\bar{x}_i | y_i = c) p(y_i = c)$$

$p(\bar{x}_i, y_i = c)$

nalysis

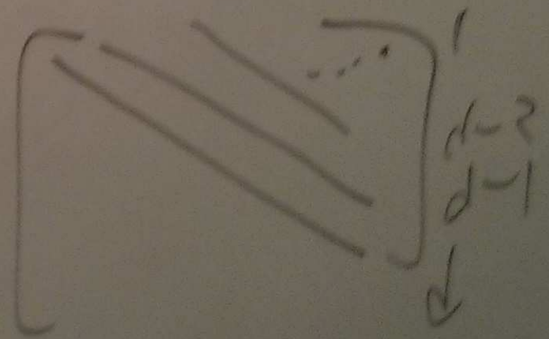
(y_i)

(μ_c, Σ_c)

parameters

$cd + c\frac{(d+1)d}{2} = O(cd^2)$

μ

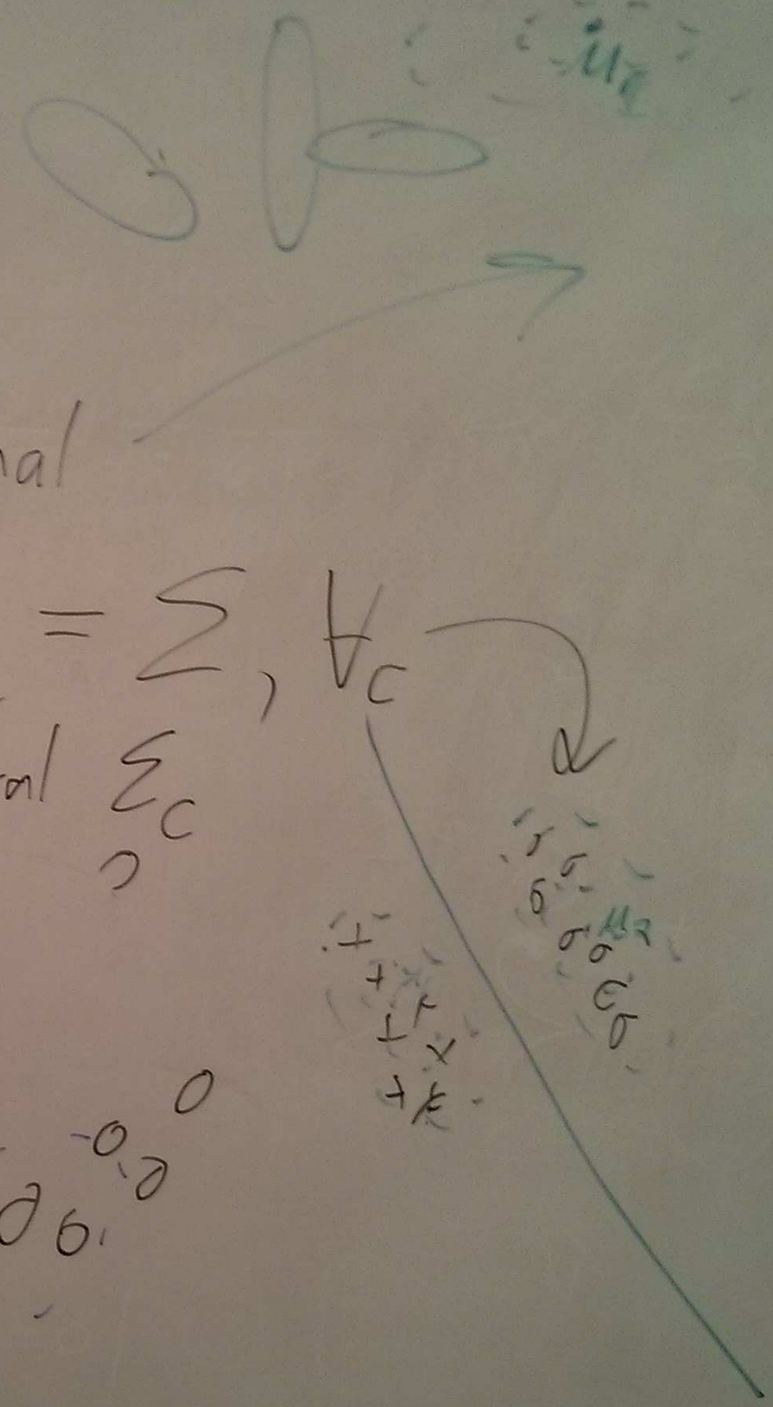


$O(Cd^2)$

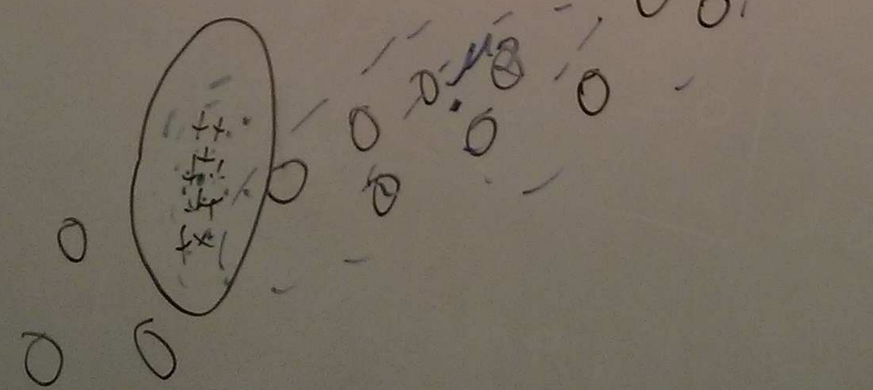
Naive Bayes: Σ_c diagonal

Linear Discrm Analysis: $\Sigma_c = \Sigma, \forall_c$

Quad Disc. Anal.: general Σ_c



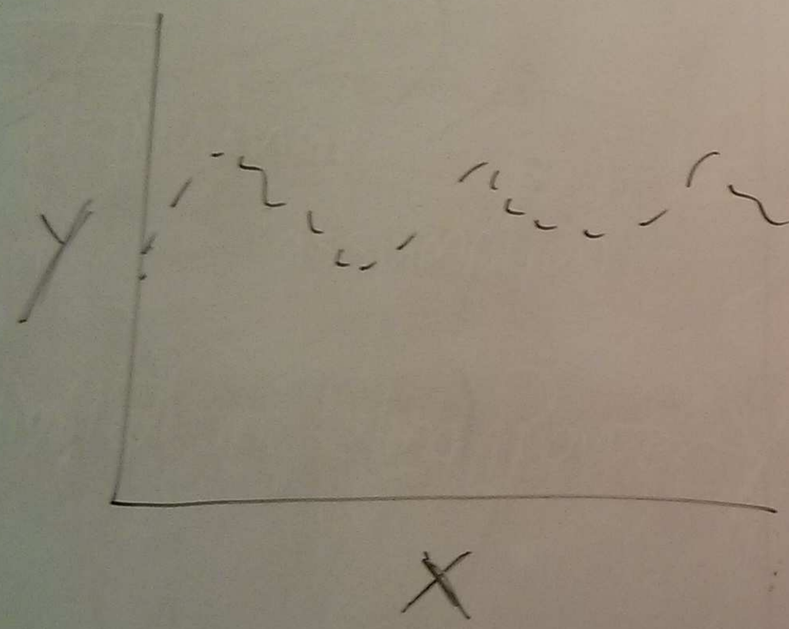
QDA:



Gaussian Regression

$$y_i = w_1 x_i^1 + w_2 x_i^2 + \dots$$

"linear model" "linear regression"



$$y_i = \bar{w}^T \bar{x}_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, 1)$$

$$y_i | \bar{x}_i, w \sim N(w^T \bar{x}_i, 1)$$

$$\underline{X} = \begin{bmatrix} - \\ \\ \end{bmatrix} \quad Y = \begin{bmatrix} | \\ y_i \\ | \end{bmatrix}$$

$$w_{MLE} = \underset{\bar{w} \in \mathbb{R}^d}{\operatorname{argmin}} -\sum_{i=1}^{N} \log p(y_i | \bar{x}_i, \bar{w})$$

$$= \underset{\bar{w} \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} (y_i - \bar{w}^T \bar{x}_i)^2$$

"Least Squares"

$$X_w = \begin{bmatrix} \bar{w}^T \bar{x}_1 \\ \bar{w}^T \bar{x}_2 \\ \vdots \\ \bar{w}^T \bar{x}_N \end{bmatrix}$$

$$NLL(w) = \underbrace{\frac{1}{2}(Y - Xw)^T(Y - Xw)}$$

$$\arg\min_{w \in \mathbb{R}^d} \frac{1}{2}(Y - Xw)^T(Y - Xw)$$

+ cons.

scalar

$$NLL(w) = \frac{1}{2}Y^TY - w^TX^Ty + \frac{1}{2}w^TX^TXw$$

d-dimensional vector

$$\nabla NLL(w) = -X^Ty + X^TXw$$

stationary

$$= 0$$

$$X^TXw = X^Ty$$

d-dimensional matrix

$$\nabla^2 NLL(w) = X^TX$$