INTRODUCTION TO MACHINE LEARNING COMPSCI 4ML3

Lecture 18

Hassan Ashtiani

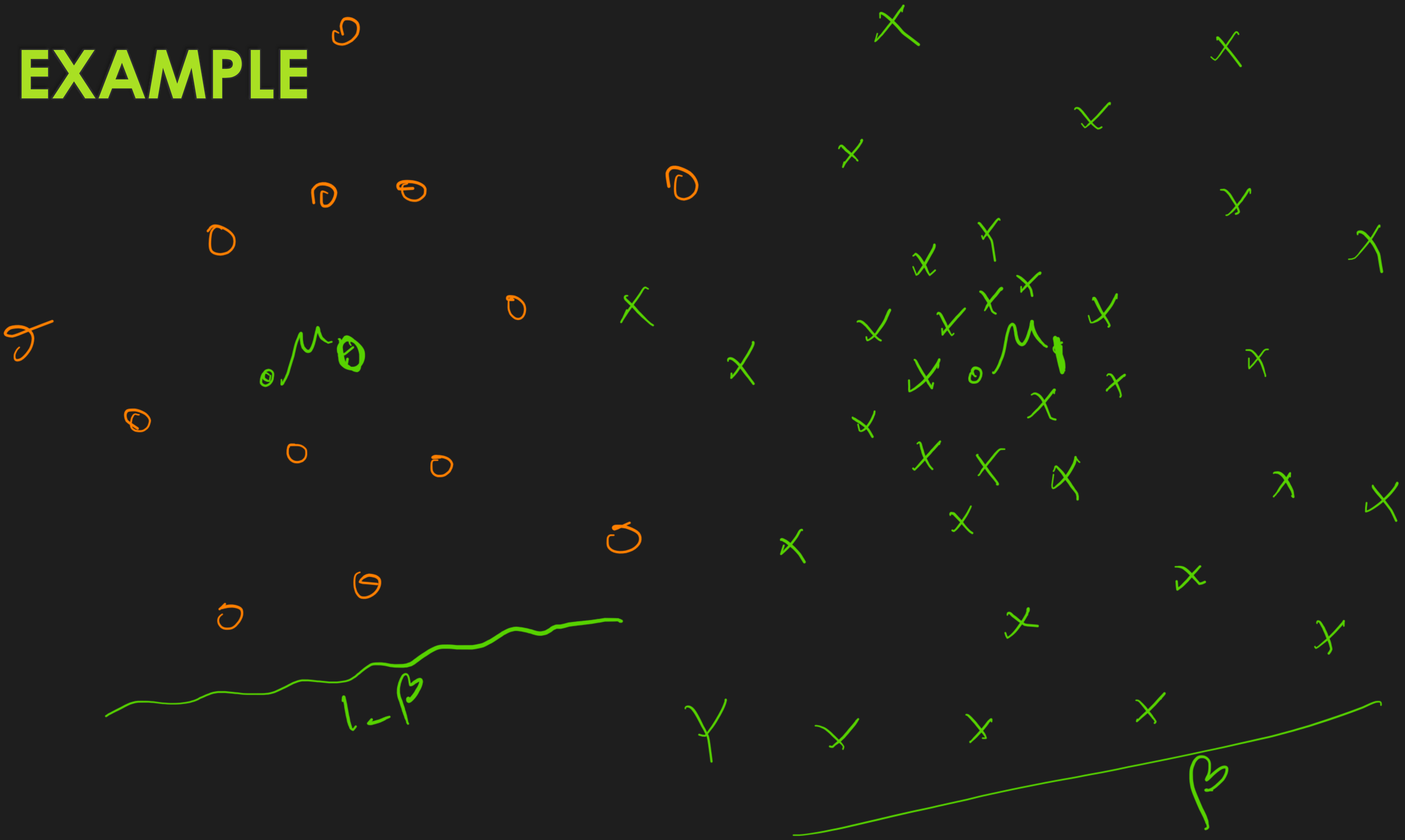
PROBABILISTIC MODELS OF CLASSIFICATION

- Assume that the data is being generated from a distribution from a certain family of (parametric) distributions
  - E.g., assume each class is a Gaussian…
- Estimate the parameters using the observed data
  - E.g., Maximum likelihood
- **Infer** the label of each test point using the estimated parameters

GAUSSIAN DISCRIMINANT ANALYSIS

- $P(x|y=0,\mu_0,\mu_1,\beta) = \frac{1}{a_0} e^{-||x-\mu_0||_2^2}$
- $P(x|y=1,\mu_0,\mu_1,\beta) = \frac{1}{a_0} e^{-||x-\mu_1||_2^2}$
- $P(y=1|\mu_0,\mu_1,\beta) = \beta$, $P(y=0|\mu_0,\mu_1,\beta) = 1-\beta$

EXAMPLE



$$\theta = \{\mu_1, \mu_2, \beta\}$$

MAXIMUM LIKELIHOOD ESTIMATE

$$\arg\max_{\theta} \; p(Z \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} p(x^i, y^i \mid \theta)$$

$= \text{------}$

Excercise --- show

Z_1: data points with label 1

Z_0: " " "

$$\mu_1 = \frac{1}{|Z_1|} \sum_{(x,y) \in Z_1} x^i ,$$

$$\mu_0 = \frac{1}{|Z_0|} \sum_{(x,y) \in Z_0} x^i , \qquad \beta = \frac{|Z_1|}{|Z_1| + |Z_0|}$$

INFERENCE

- Now we have estimated the parameters μ_0, μ_1, β
- How can we predict the label of a new test point?

INFERENCE

$$p(y=0 \mid x, \theta) \overset{?}{>} p(y=1 \mid x, \theta)$$

So check

$$\frac{p(y=0 \mid x, \theta)}{p(y=1 \mid x, \theta)} \geq 1$$

$$\Longleftrightarrow \frac{p(y=0, x, \theta)\, p(x, \theta)}{p(x, \theta)\, p(y=1, x, \theta)} \geq 1$$

$$\frac{p(y=0, x, \theta)}{p(y=1, x, \theta)} \overset{?}{\geq} 1$$

$$\Longleftrightarrow \frac{p(x \mid y=0, \theta)\, p(y=0)\, p(\theta)}{p(x \mid y=1, \theta)\, p(y=1)\, p(\theta)} \geq 1$$

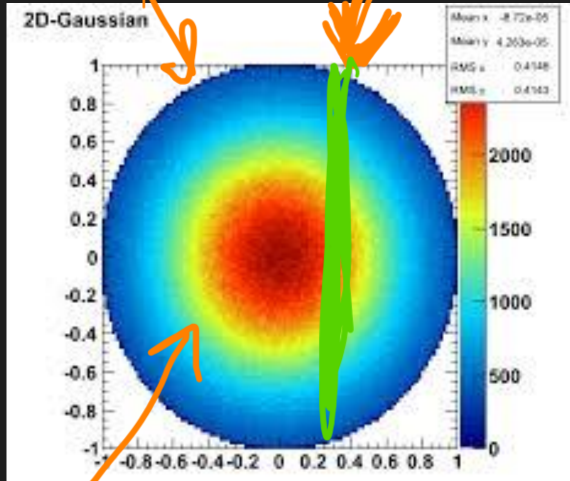
$$\Longleftrightarrow \frac{e^{-\|x-\mu_0\|^2} \cdot (1-\beta)}{e^{-\|x-\mu_1\|^2} \cdot \beta} \geq 1$$

COMPUTING $P(Y=0 \mid X, \ldots)$?

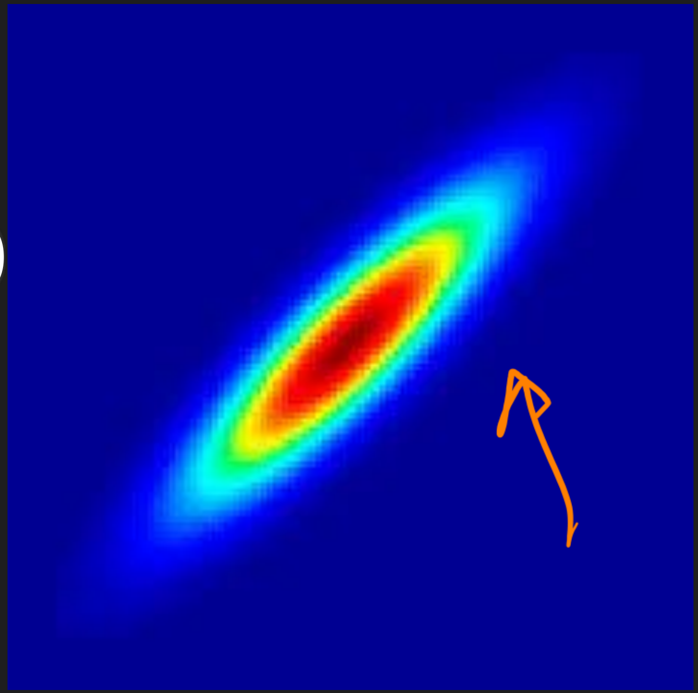
GENERALIZATION: CORRELATED FEATURES

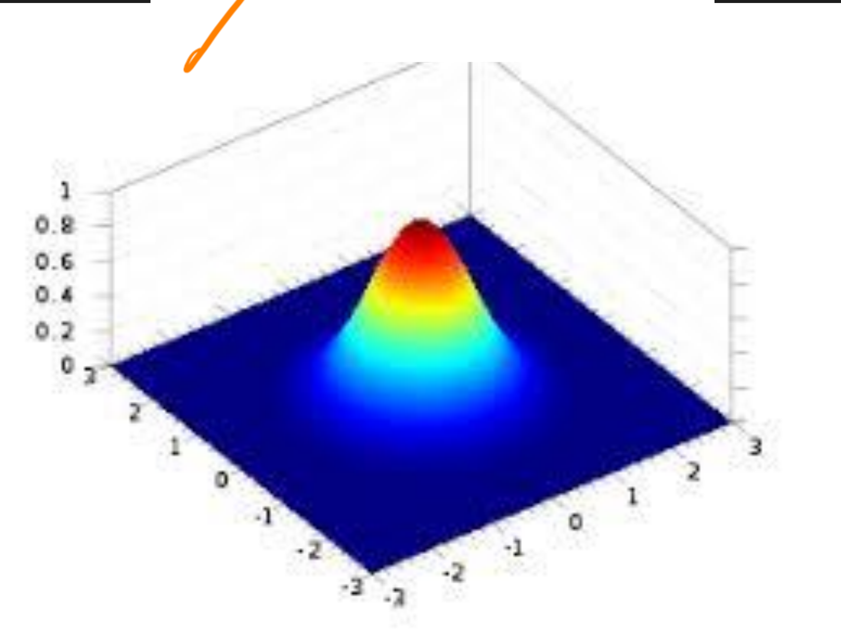
- Gaussian with correlated features

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



- Covariance matrix:
- $\Sigma = E[(x-\mu)(x-\mu)$





NAÏVE BAYES CLASSIFIERS

- The Naïve Bayes assumption:
  - Given the label, the coordinates are statistically independent
  - $P(x|y = k, \Theta) = \pi_j P(x_j|y = k, \Theta)$
- Choices for $P(x|y = i, \Theta)$
  - Gaussian, categorical, binomial, etc.
- How to find most probable y?

Gaussian Discriminat Analysis

NAÏVE BAYES – INFERENCE

Naïve Bayes Classifiers

- Only need to estimate the distribution of each coordinate separately given the label.
  - No curse of dimensionality
  - Fast computation for learning and prediction
  - Assumption is very strong – may be far from reality

GENERATIVE (GAUSSIAN) ASSUMPTION

- Too restrictive?

GENERATIVE VS DISCRIMINATIVE MODELS

- Probabilistic Generative models
  - Try to model $P(x, y)$
  - E.g., model $P(x|y)$ and $P(y)$
  - We are modeling $P(x)$ as well
- Probabilistic Discriminative models
  - Try to model $P(y|x)$ only
  - No need to know or model $P(x)$
  - E.g., logistic regression
- Non-probabilistic discriminative Approaches