Classification – Bayes optimal classifier

Aarti Singh

Machine Learning 10-315
Aug 28, 2019

Logistics

- Add yourself to 10-315 on Piazza
- Recitation on Friday – Probability review
- Office hours
  - Mon Siddharth 1-2 pm
  - Tues Yue TBA
  - Wed Aarti 9:30-10:30 am outside classroom
  - Thurs Fabricio 11 am-12 noon

- QnA1 to be released TODAY on Canvas

Performance Measure

For a random test data X, measure of closeness between true label *Y* and prediction *f*(*X*)

Binary Classification $\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}}$ **0/1 loss**

Regression $\text{loss}(Y, f(X)) = (f(X) - Y)^2$ **square loss**

- What if overestimating stock price is 10 times more costly than underestimating it?
- What if missing a tumor is 10 times more costly than falsely detecting it?

Performance Measure

For a random test data X, measure of closeness between true label *Y* and prediction *f*(*X*)

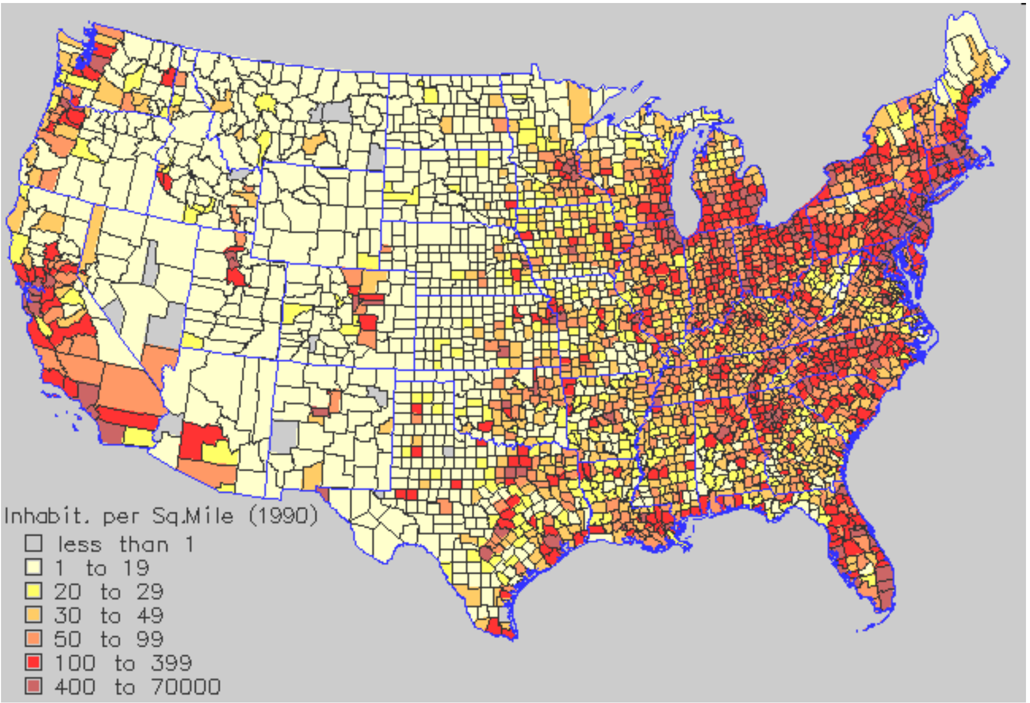
Binary Classification $\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}}$ **0/1 loss**

Regression $\text{loss}(Y, f(X)) = (f(X) - Y)^2$ **square loss**

Density Estimation?

Unsupervised Learning

Density/Distribution Estimation



Population density

Bias of a coin

Performance Measure

For a random test data X, measure of closeness between true label *Y* and prediction *f*(*X*)

Binary Classification $\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}}$ **0/1 loss**

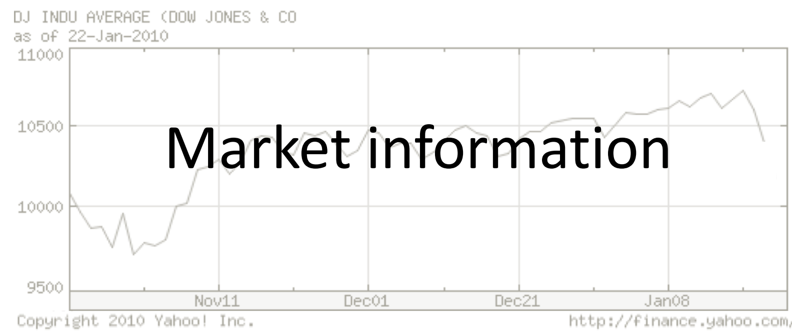
Regression $\text{loss}(Y, f(X)) = (f(X) - Y)^2$ **square loss**

Density Estimation $\text{loss}(f(X)) = -\log(\mathbb{P}_f(X))$ **Negative log likelihood loss**

Notion of "Features"

Input $X \in \mathcal{X}$

Input $X \in \mathcal{X}$



- How to represent inputs mathematically?
- Document vector X = list of words (different length for each document)

  frequency of words (length of each document = size of vocabulary)
- Market information X = daily/monthly? price of share for past 10 years
- Image X = intensity at each pixel, fourier transform values, SIFT etc.

Classification

Goal: Construct **prediction rule** $f : \mathcal{X} \rightarrow \mathcal{Y}$

Sports
Science
News

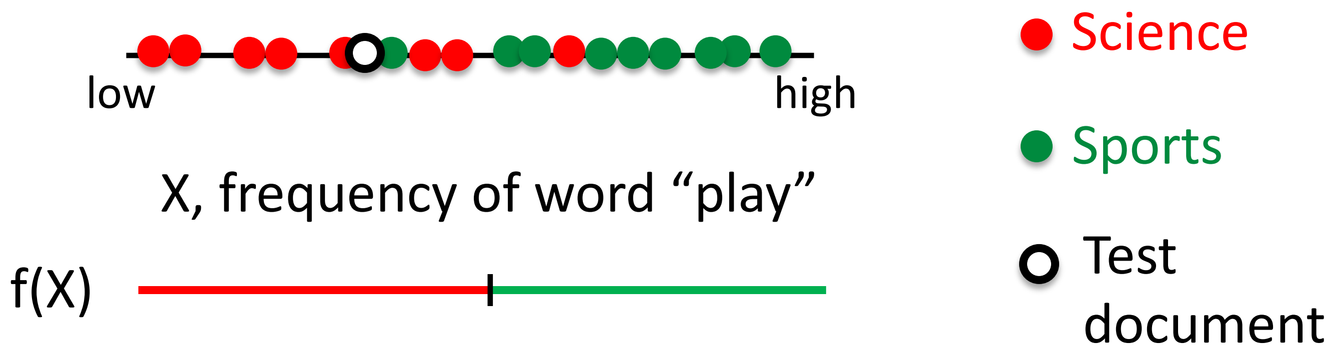
Input feature vector, X

Label, Y

In general: label Y can belong to more than two classes
X is multi-dimensional (many features represent an input)

But lets start with a simple case:
label Y is binary (either "Sports" or "Science")
X is frequency of word "play" = count/total length of document

Binary Classification



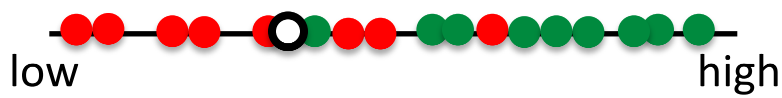
Model X and Y as random variables with joint distribution P_{XY}

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

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

Training and test data are independent draws from **same** distribution

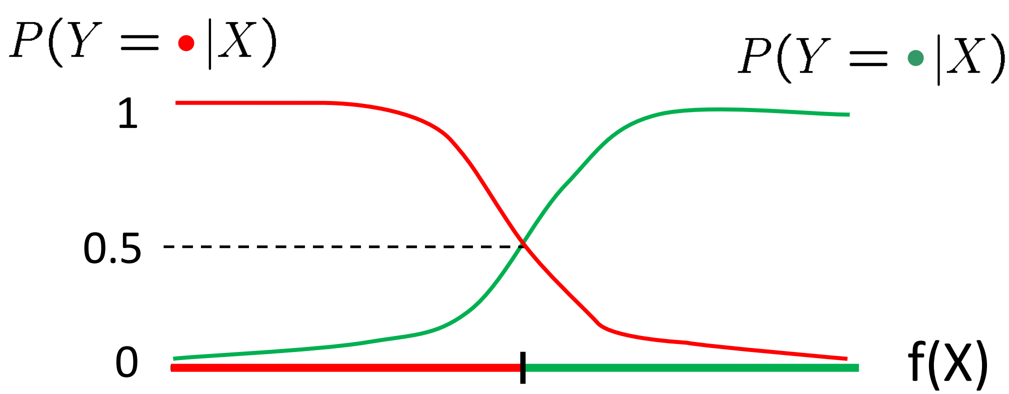
Binary Classification



- Science
- Sports
- Test document

X, frequency of word “play”

Model X and Y as random variables



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

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

Optimal Classifier

Optimal classifier: $f^*(x) = \arg\max_{Y=y} P(Y = y|X = x)$

Why??

Goal: Construct **prediction rule** $f^* : \mathcal{X} \rightarrow \mathcal{Y}$ that minimizes $\text{loss}(Y, f(X))$ for a randomly drawn test data (X,Y)

$$\min_f \mathbb{E}_{XY}\left[\text{loss}(Y, f(X))\right]$$

$$= \min_f \mathbb{E}_{XY}[\mathbf{1}_{\{f(X) \neq Y\}}] \quad \textbf{0/1 loss}$$

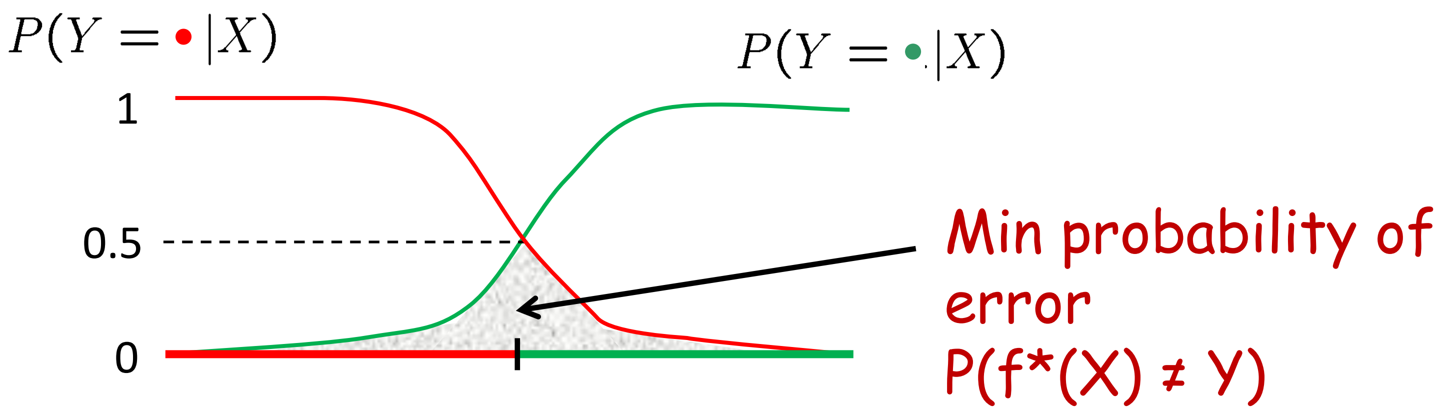
$$= \min_f \mathbb{P}_{XY}(f(X) \neq Y) \quad \textbf{Probability of Error}$$

Minimizer is indeed f*!!

HW1!

Error of Optimal Classifier

Optimal classifier: $f^*(x) = \arg\max_{Y=y} P(Y = y|X = x)$



- Even the optimal classifier makes mistakes: min probability of error > 0

Bayes Optimal Classifier

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 Optimal Classifier

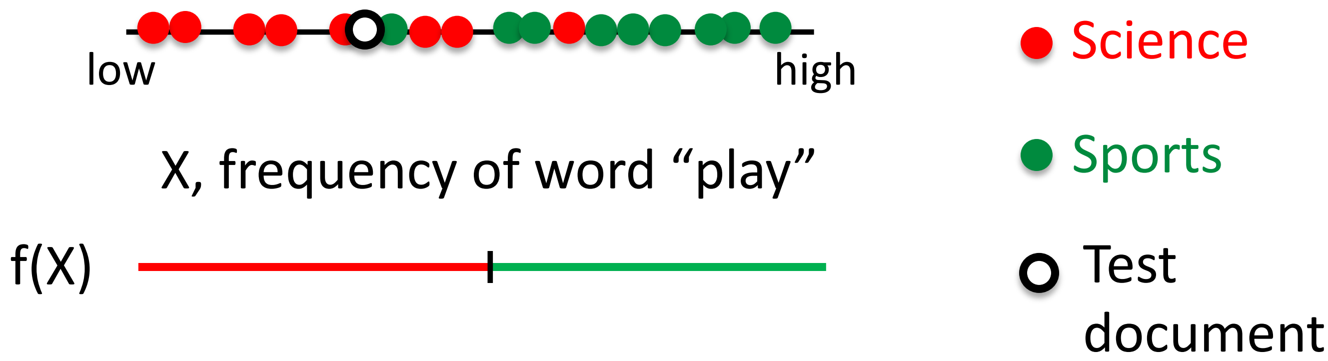
Bayes Rule: $$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(Y = y|X = x) \;=\; \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

Bayes Optimal classifier:

$$\begin{aligned} f^*(x) &= \arg\max_{Y=y} P(Y = y|X = x) \\ &= \arg\max_{Y=y} \underbrace{P(X = x|Y = y)}_{\text{Class conditional distribution}}\underbrace{P(Y = y)}_{\text{Class probability}} \end{aligned}$$

Bayes Optimal Classifier



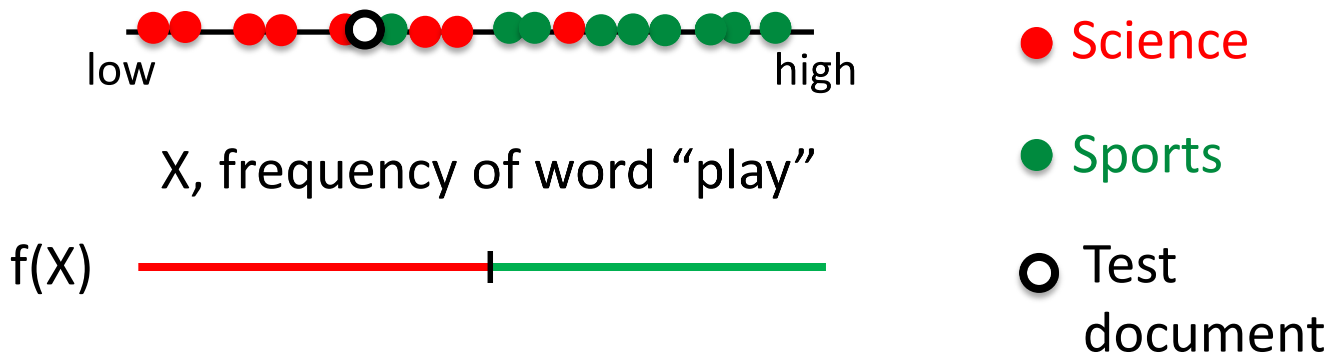
$$f^*(x) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional distribution}} \underbrace{P(Y=y)}_{\text{Class probability}}$$

We can now consider appropriate models for the two terms:

Class probability P(Y=y)

Class conditional distribution of features P(X=x|Y=y)

Modeling class probability

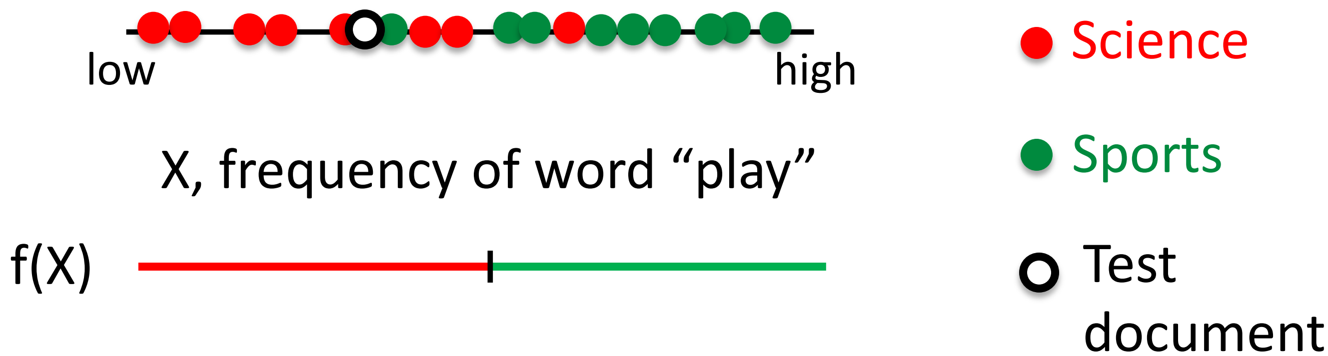


Modeling Class probability P(Y=y) = Bernoulli(θ)

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

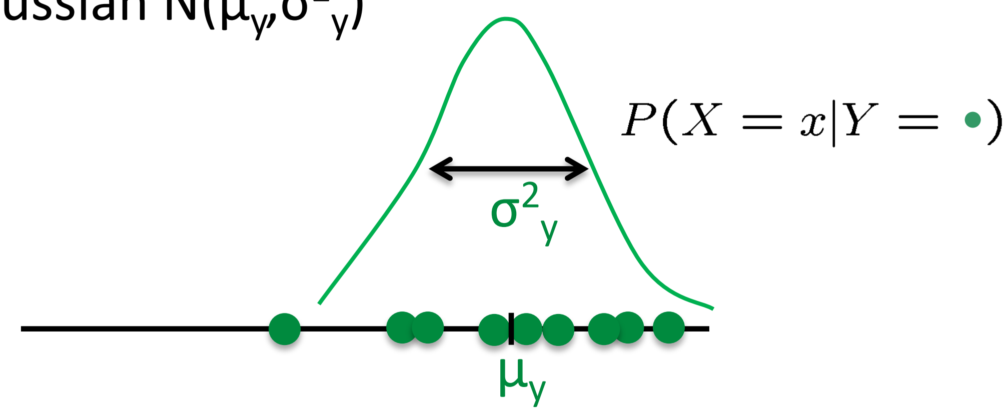
Like a coin flip

Modeling class conditional distribution of features



Modeling Class Conditional distribution of features P(X=x|Y=y)

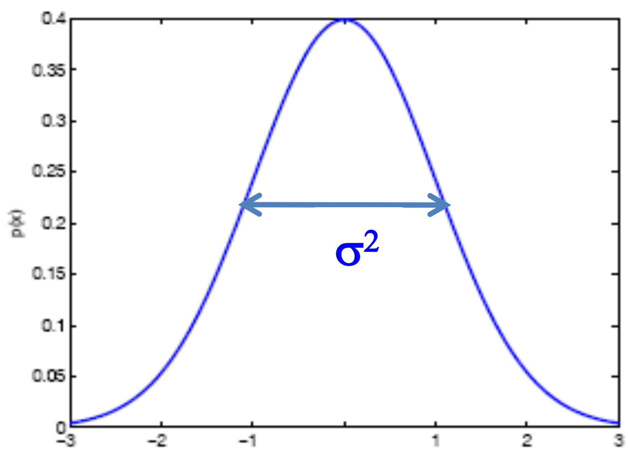
E.g. P(X=x|Y=y) = Gaussian N(μ_y,σ^2_y)



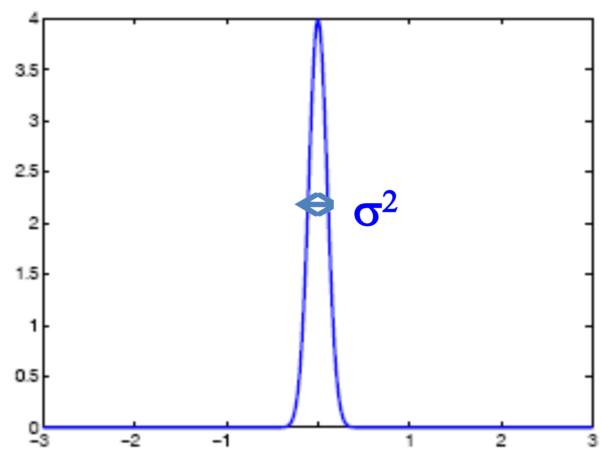
1-dim Gaussian distribution

X is Gaussian N(μ,σ²)

$$P(X = x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



μ=0



μ=0

d-dim Gaussian distribution

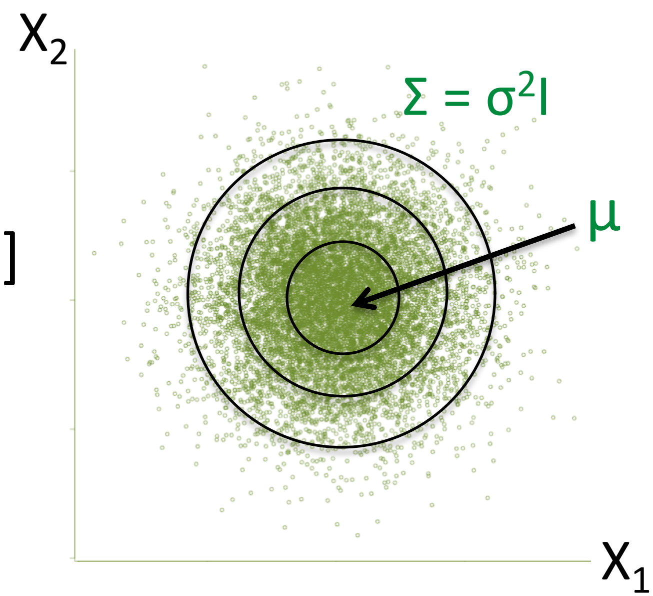
X is Gaussian N(μ, Σ)

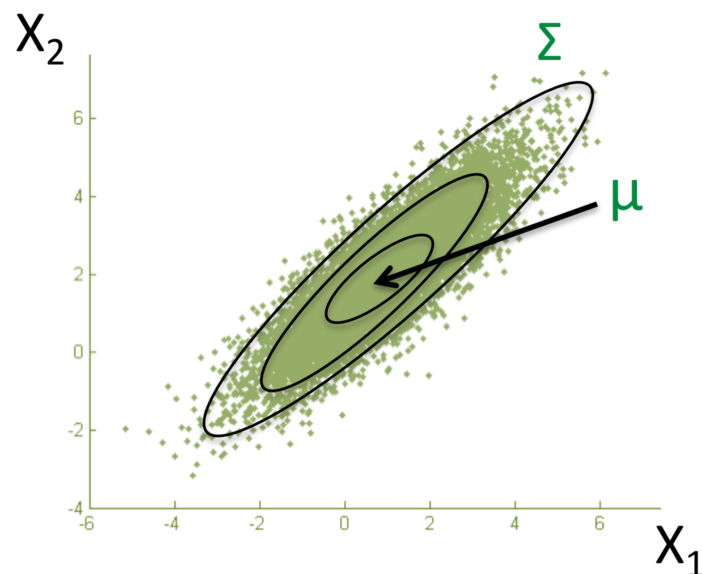
μ is d-dim vector, Σ is dxd dim matrix

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

d=2

$X = [X_1; X_2]$





Gaussian Bayes classifier

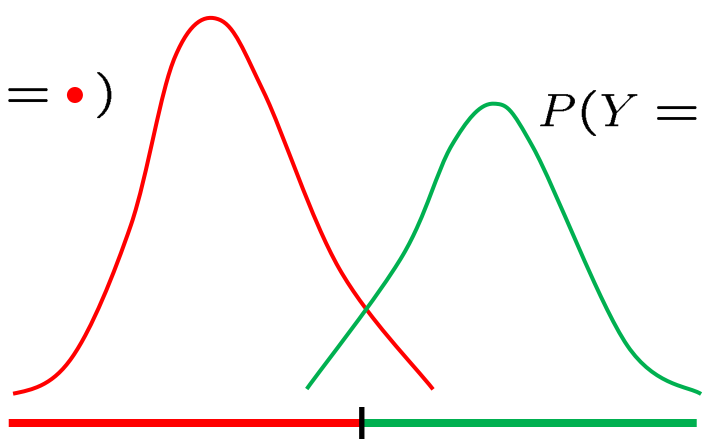
$$f^*(x) = \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{Class conditional density}} \underbrace{P(Y=y)}_{\text{Class probability}}$$

How to learn parameters θ, μ_y, Σ_y from data?

Gaussian(μ_y,Σ_y)

Bernoulli(θ)

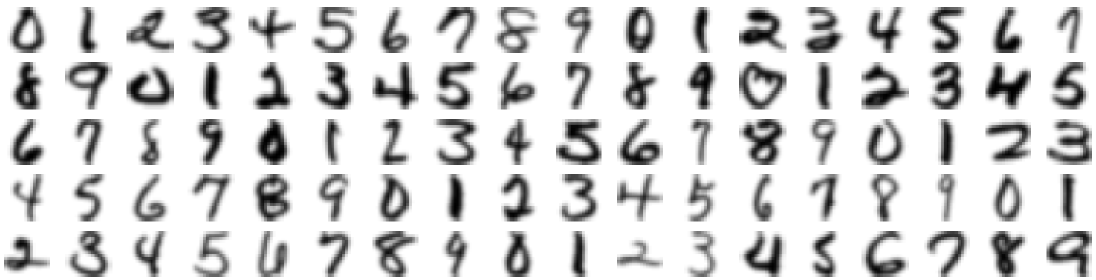
$P(Y = \bullet)P(X = x|Y = \bullet)$



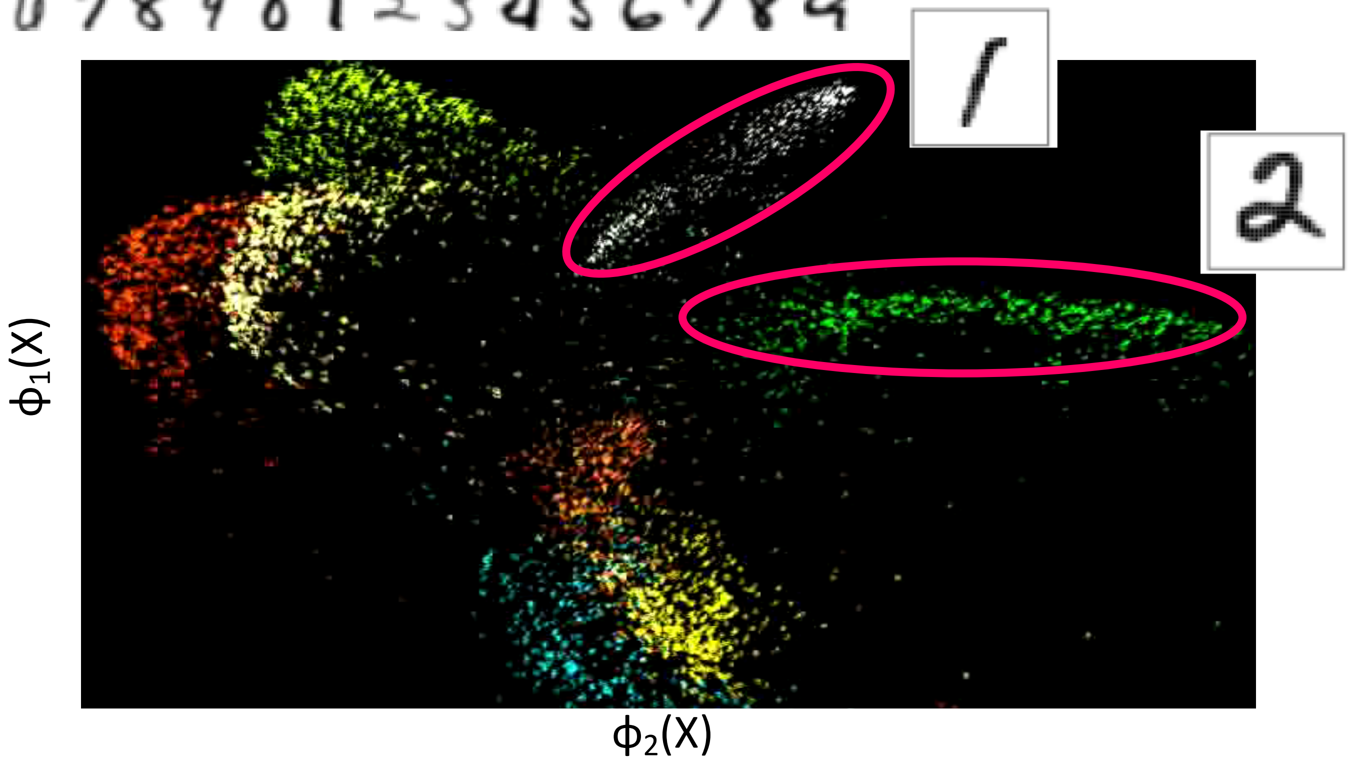
$P(Y = \bullet)P(X = x|Y = \bullet)$

Multi-class problem
Multi-dimensional input X

Handwritten digit recognition



Multi-class classification



Note: 8 digits shown out of 10 (0, 1, …, 9);
Axes are obtained by nonlinear dimensionality reduction (later in course)

Handwritten digit recognition

Each image represented as a vector of **intensity values at the d pixels (features)**

Training Data:

Input, X   ... n greyscale images

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_d \end{bmatrix}$$

Label, Y 1 2 ... n labels

Gaussian Bayes model:

$P(Y = y) = p_y$ for all y in 0, 1, 2, ..., 9

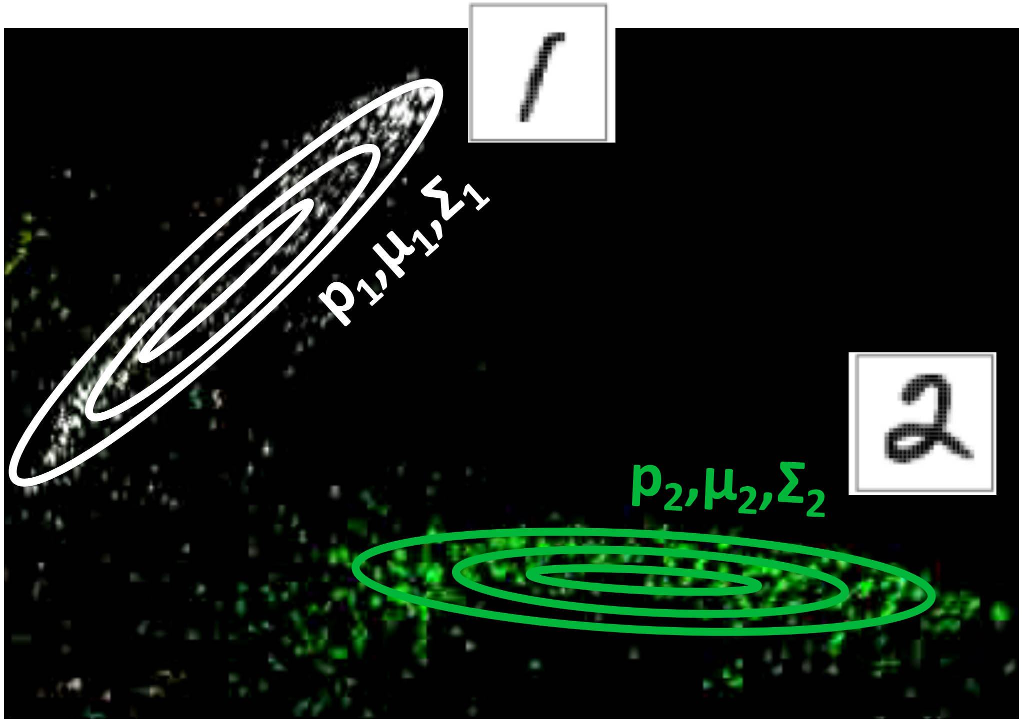
$p_0, p_1, \ldots, p_9$ (sum to 1)

$P(X=x|Y = y) \sim N(\mu_y,\Sigma_y)$ for each y

μ_y – d-dim vector

Σ_y - dxd matrix

Gaussian Bayes classifier



How to learn parameters p_y, μ_y, Σ_y from data?

$P(Y = y) = p_y$ for all y in 0, 1, 2, …, 9

$P(X=x|Y = y) \sim N(\mu_y,\Sigma_y)$ for each y

$p_0, p_1, \ldots, p_9$ (sum to 1)

μ_y – d-dim vector

Σ_y - dxd matrix

How many parameters do we need to learn?

Class probability:

$P(Y = y) = p_y$ for all y in 0, 1, 2, …, 9 $\qquad$ $p_0, p_1, \ldots, p_9$ (sum to 1)

K-1 if K labels

Class conditional distribution of features:

$P(X=x|Y = y) \sim N(\mu_y,\Sigma_y)$ for each y $\qquad$ μ_y – d-dim vector

Σ_y - dxd matrix

$Kd + Kd(d+1)/2 = O(Kd^2)$ if d features

Quadratic in dimension d! If d = 256x256 pixels, ~ 21.5 billion parameters!