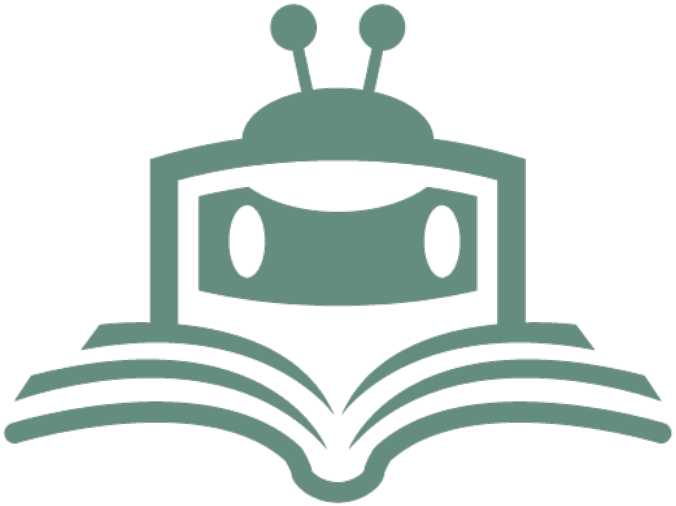


MATHEMATICS FOR MACHINE LEARNING

Nagoya University, Fall 2022

Lecture 7

Naive Bayes II & Gaussian discriminant analysis

https://www.henrikbachmann.com/mml2022.html

Recall

Generative vs Discriminative learning algorithm

Logistic regression

x: feature (e.g. hours of studying)
y: label (e.g. passing or failing exam)

Want to find a hypothesis which describes P(y|x)

$$P(y = 1 \mid x; \theta) = h_\theta(x)\,.$$

Learning P(y|x) is an example of a **discriminative learning algorithm.**

Generative learning algorithm: Learn P(x|y) and P(y) and then use Bayes rule

$$P(y = 1 \mid x) = \frac{P(x \mid y = 1)P(y = 1)}{P(x)}$$

together with the **Naive Bayes assumption** that

$$P(x \mid y) = P(x_1, \ldots, x_d|y) = \prod_{i=1}^{d} P(x_i \mid y)$$

and the **law of total probability**

$$P(x) = P(x \mid y = 1)P(y = 1) + P(x \mid y = 0)P(y = 0)$$

Recall

Email spam filter

Let's assume we want to create an email spam filter by using supervised learning.

Features: Email

$\mathcal{X} =$ Emails

Labels: No Spam & Spam

$\mathcal{Y} = \{0, 1\}$

Training set

	Email text	No spam / Spam
1	Do math today	0
2	Buy	1
3	Buy book	0
4	Today do math drugs	1
5	Buy drugs book today	1

According to this training set: Is the email “Buy book today” spam?

Recall

Email spam filter: From emails to vectors

Training set

	Email text	No spam / Spam
1	Do math today	0
2	Buy	1
3	Buy book	0
4	Today do math drugs	1
5	Buy drugs book today	1

Create a dictionary and assign to an email a vector

1	book
2	buy
3	do
4	drugs
5	math
6	today

"Buy book today" $\longleftrightarrow$ $\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \in \mathcal{X} = \{0,1\}^6$

We will ignore the order and the number of appearances of words

Recall

Email spam filter

Training set

	Email text	No spam / Spam
1	Do math today	0
2	Buy	1
3	Buy book	0
4	Today do math drugs	1
5	Buy drugs book today	1

The probabilities according to our training set:

i		appears in non-spam: $P(x_i = 1 \mid y = 0)$	appears in spam: $P(x_i = 1 \mid y = 1)$
1	book	$\frac{1}{2}$	$\frac{1}{3}$
2	buy	$\frac{1}{2}$	$\frac{2}{3}$
3	do	$\frac{1}{2}$	$\frac{1}{3}$
4	drugs	0	$\frac{2}{3}$
5	math	$\frac{1}{2}$	$\frac{1}{3}$
6	today	$\frac{1}{2}$	$\frac{2}{3}$

$$P(y = 0) = \frac{2}{5} \qquad P(y = 1) = \frac{3}{5}$$

Recall

Email spam filter

$$P(y=0)=\frac{2}{5} \qquad P(y=1)=\frac{3}{5}$$

$$x=\begin{pmatrix}1\\1\\0\\0\\0\\1\end{pmatrix}$$

i		appears in non-spam: $P(x_i=1 \mid y=0)$	appears in spam: $P(x_i=1 \mid y=1)$
1	book	$\frac{1}{2}$	$\frac{1}{3}$
2	buy	$\frac{1}{2}$	$\frac{2}{3}$
3	do	$\frac{1}{2}$	$\frac{1}{3}$
4	drugs	0	$\frac{2}{3}$
5	math	$\frac{1}{2}$	$\frac{1}{3}$
6	today	$\frac{1}{2}$	$\frac{2}{3}$

"Buy book today"

$$\prod_{i=1}^{d} P(x_i \mid y=0)=\frac{1}{2}\cdot\frac{1}{2}\cdot\left(1-\frac{1}{2}\right)\cdot(1-0)\cdot\left(1-\frac{1}{2}\right)\cdot\frac{1}{2}=\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot 1\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{32},$$

$$\prod_{i=1}^{d} P(x_i \mid y=1)=\frac{1}{3}\cdot\frac{2}{3}\cdot\left(1-\frac{1}{3}\right)\cdot\left(1-\frac{2}{3}\right)\cdot\left(1-\frac{1}{3}\right)\cdot\frac{2}{3}=\frac{1}{3}\cdot\frac{2}{3}\cdot\frac{2}{3}\cdot\frac{1}{3}\cdot\frac{2}{3}\cdot\frac{2}{3}=\frac{16}{729}.$$

$$P(y=1 \mid x)=\frac{P(x \mid y=1)P(y=1)}{P(x)}$$

$$=\frac{\prod_{i=1}^{d} P(x_i \mid y=1)\cdot\phi_{y=1}}{\prod_{i=1}^{d} P(x_i \mid y=1)\cdot\phi_{y=1}+\prod_{i=1}^{d} P(x_i \mid y=0)(1-\phi_{y=1})}=\frac{\frac{16}{729}\cdot\frac{3}{5}}{\frac{16}{729}\cdot\frac{3}{5}+\frac{1}{32}\cdot\frac{2}{5}}=\frac{256}{499}\approx 0.51$$

Recall

Naive Bayes assumption

Naive Bayes assumption:

The features are "conditionally independent" given the label.

conditionally independent

A and B are conditionally independent given C if and only if, given knowledge that C occurs, knowledge of whether A occurs provides no information on the likelihood of B occurring, and knowledge of whether B occurs provides no information on the likelihood of A occurring.

Recall

Conditionally independent: Example

A: Ability to do math of a person

B: Foot size of a person

A and B are **not independent**, since if I tell you someone's foot size, it hints at their age, which in turn hints at their ability to do math.

C: Age of a person

A and B are **conditionally independent given C**, since if tell you someone's age (C), then the ability of doing math (A) will not change whether you know the foot size (B) or not, i.e.

$$P(A \mid C) = P(A \mid B,C)$$

Recall

Naive Bayes classifier: Training

$$\phi_{i|y=1} = P(x_i = 1 \mid y = 1)\,,$$
$$\phi_{i|y=0} = P(x_i = 1 \mid y = 0)\,,$$
$$\phi_{y=1} = P(y = 1)\,.$$

Indicator function

$$I(S) = \begin{cases} 1\,, & S \text{ is true} \\ 0\,, & S \text{ is false} \end{cases}.$$

Given a training set $\mathcal{T} = \big((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\big)$ we can calculate them by

$$\phi_{i|y=1} = \frac{\sum_{j=1}^{n} I(x_i^{(j)} = 1 \wedge y^{(j)} = 1)}{\sum_{j=1}^{n} I(y^{(j)} = 1)}$$

$$\phi_{i|y=0} = \frac{\sum_{j=1}^{n} I(x_i^{(j)} = 1 \wedge y^{(j)} = 0)}{\sum_{j=1}^{n} I(y^{(j)} = 0)}$$

$$\phi_{y=1} = \frac{1}{n} \sum_{j=1}^{n} I(y^{(j)} = 1)$$

Recall

Naive Bayes classifier: Judging a new email

Training: Determine the values

$$\begin{aligned}\phi_{i|y=1} &= P(x_i = 1 \mid y = 1)\,,\\ \phi_{i|y=0} &= P(x_i = 1 \mid y = 0)\,,\\ \phi_{y=1} &= P(y = 1)\,.\end{aligned}$$

The probability of a new email $x = (x_1, \ldots, x_d)^T$ being spam is then

$$\begin{aligned}P(y = 1 \mid x) &= \frac{P(x \mid y = 1)P(y = 1)}{P(x)}\\ &= \frac{\prod_{i=1}^{d} P(x_i \mid y = 1) \cdot \phi_{y=1}}{\prod_{i=1}^{d} P(x_i \mid y = 1) \cdot \phi_{y=1} + \prod_{i=1}^{d} P(x_i \mid y = 0)(1 - \phi_{y=1})}\,.\end{aligned}$$

$$\begin{aligned}P(x_i = 1 \mid y = 1) &= \phi_{i|y=1}\,,\\ P(x_i = 0 \mid y = 1) &= 1 - \phi_{i|y=1}\,,\end{aligned}$$

$$\begin{aligned}P(x_i = 1 \mid y = 0) &= \phi_{i|y=0}\,,\\ P(x_i = 0 \mid y = 0) &= 1 - \phi_{i|y=0}\,.\end{aligned}$$

Naive Bayes classifier: Laplace smoothing

We should never assume an event to have probability 0 or 1.

For this we will use **Laplace smoothing** and change our parameters as follows:

$$\widetilde{\phi}_{i|y=1} = \frac{1 + \sum_{j=1}^{n} I(x_i^{(j)} = 1 \wedge y^{(j)} = 1)}{2 + \sum_{j=1}^{n} I(y^{(j)} = 1)}$$

$$\widetilde{\phi}_{i|y=0} = \frac{1 + \sum_{j=1}^{n} I(x_i^{(j)} = 1 \wedge y^{(j)} = 0)}{2 + \sum_{j=1}^{n} I(y^{(j)} = 0)}$$

Interpretation: Add 4 imaginary emails

- A spam email which contains every word.
- A spam email which does not contain any word.
- A non-spam email which contains every word.
- A non-spam email which does not contain any word.

Naive Bayes classifier: Laplace smoothing

https://colab.research.google.com/drive/1XU8NcIWbf1io_dDbMKLKqnKuy9MmNqu2?usp=sharing

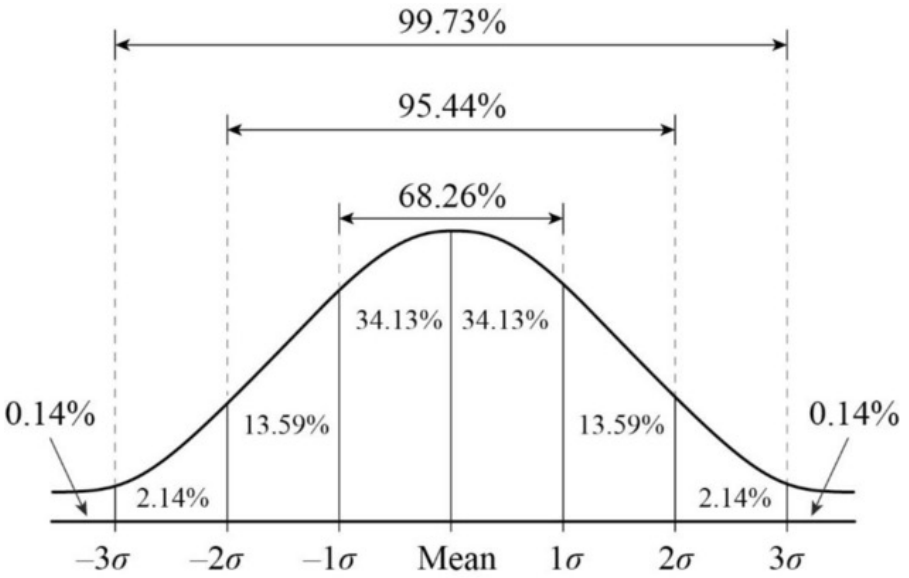
Normal/Gaussian distribution

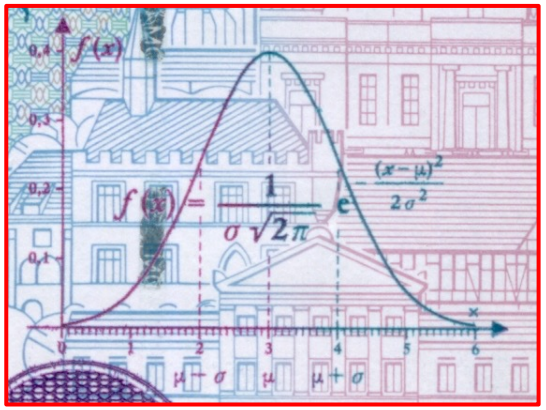
probability density function

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

$\mu \in \mathbb{R}$ is the **mean**

$\sigma \in \mathbb{R}$ is the **standard deviation**





A lot of things follow a normal distribution, e.g. height of people, exam results, etc.

Multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$

probability density function

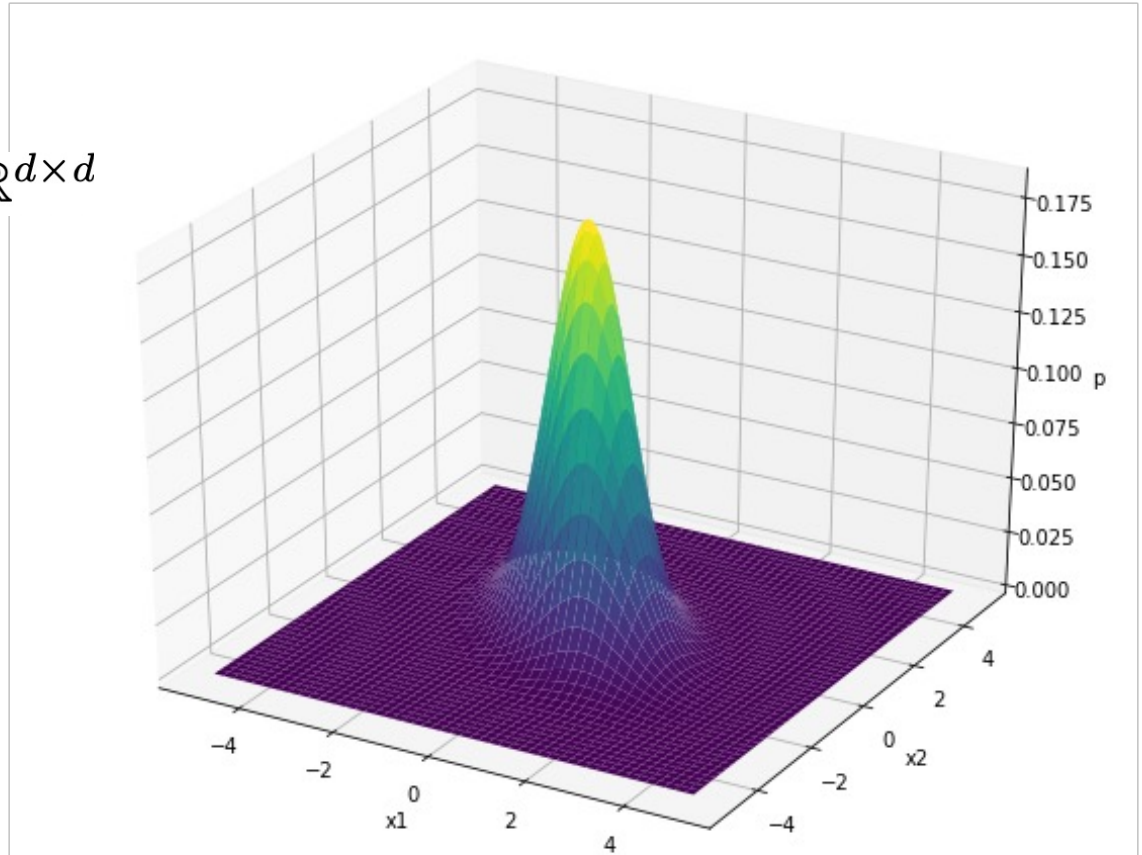
$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

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

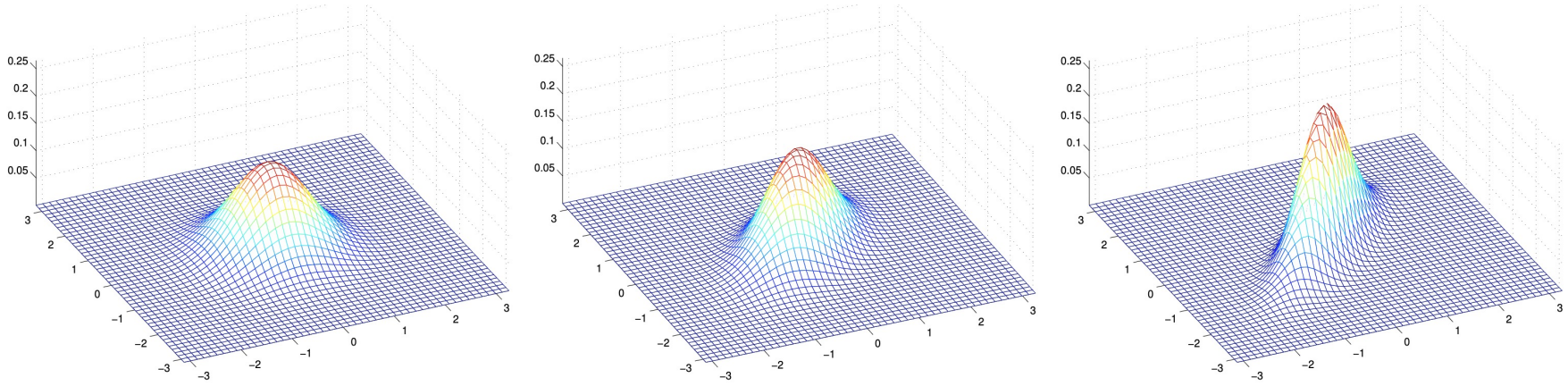
covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$

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

”x follows a multivariate normal distribution”



Multivariate normal distribution



The figures above show Gaussians with mean 0, and with covariance matrices respectively

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

Gaussian discriminant analysis (GDA) model

For the **Gaussian discriminant analysis (GDA)** model we assume that we have d features, i.e. $\mathcal{X} = \mathbb{R}^d$, and again two labels $\mathcal{Y} = \{0, 1\}$. We want to model again $p(x \mid y)$ by assuming that

$$\begin{aligned} y &\sim \text{Bernoulli}(\phi)\,, \\ x \mid y = 0 &\sim \mathcal{N}(\mu_0, \Sigma)\,, \\ x \mid y = 1 &\sim \mathcal{N}(\mu_1, \Sigma) \end{aligned}$$

for some $\mu_0, \mu_1 \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ and $\phi \in \mathbb{R}$, i.e. we have

$$p(x|y=0) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)\,,$$

$$p(x|y=1) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)\,.$$

Here $y \sim \text{Bernoulli}(\phi)$ means y follows a Bernoulli distribution, which just means that $p(y=1) = \phi$ and $p(y=0) = 1-\phi$, which we can again write compactly as

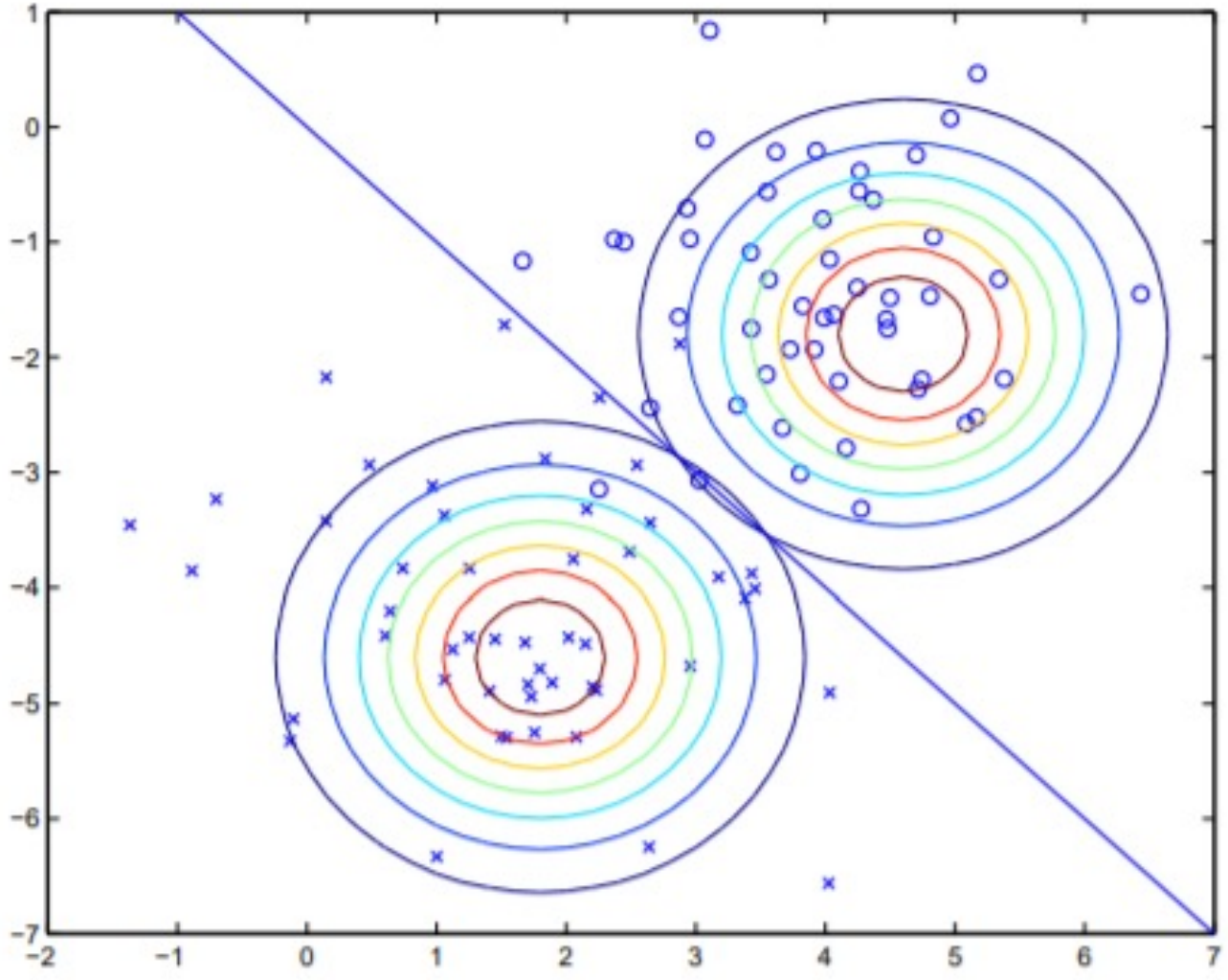
$$p(y) = \phi^y(1-\phi)^{1-y}\,.$$

Goal: Given a training set, find the best possible parameters

$$\mu_0, \mu_1 \in \mathbb{R}^d,\ \Sigma \in \mathbb{R}^{d \times d} \text{ and } \phi \in \mathbb{R}$$

Gaussian discriminant analysis (GDA) model

d=2



Gaussian discriminant analysis (GDA) model

Proposition 3.14. *The log-likelihood gets maximized by choosing the following parameters*

$$\phi = \frac{1}{n}\sum_{i=1}^{n} I(y^{(i)} = 1)\,,$$

$$\mu_c = \frac{\sum_{i=1}^{n} I(y^{(i)} = c)x^{(i)}}{\sum_{i=1}^{n} I(y^{(i)} = c)}, \qquad (c \in \{0, 1\})$$

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T\,.$$