Bayes Classifier with Multivariate Normal Distribution

Slides from Machine Learning by Ethem Alpaydin
Chapter 4 and 5

Expanded by some slides from Gutierrez-Osuna

Overview

- We learned how to use the Bayesian approach for classification *if* we had the probability distribution of the underlying classes ($p(\mathbf{x}|C_i)$).

- **Naive Bayes** approached the problem by simplifying the task, with the assumption of:

$$P(\mathbf{x} \mid C_i) = \prod_{j=1..d} P(x_j \mid C_i)$$

- Now we are going to look at how to use the **multivariate Normal distribution** to model $P(\mathbf{x} \mid C_i)$

Overview

- Gaussian Bayes classifier …
  - is a Bayesian classifier so it assigns the input **x** to the class C_j for which $P(C_j|\mathbf{x})$ is largest;
  - assumes that **x** is multivariate normal.
    - $\mathbf{x} \sim N(\mathbf{x} \mid \mu, \Sigma)$
    - I.e. $p(\mathbf{x} \mid C_j) = N(\mathbf{x} \mid \mu_i, \Sigma_i)$
  - It computes $p(\mathbf{x} \mid C_j)$ by estimating the parameters of the multivariate Normal distribution from the data.
    - For for 1D-Normal distribution:
      - Estimate μ, σ^2 using Maximum Likelihood Estimation
      - Use $p(\mathbf{x} \mid C_j) = N(\mathbf{x} \mid \mu_i, \Sigma_i)$ in Bayesian decision process.

Self study - Skip

Review: Expectation

- A discrete random variable X takes on values $x_1..x_m$ with probabilities $p(x_1)...p(x_m)$.

- Its expected value is defined as $$E(X) = \sum_i x_i P(x_i)$$

- E.g. In a TV game, you will lose -1000 TL with 0.6 probability; and gain 10,000 TL with 0.4 probability. What is the expected reward?

- A discrete random variable X takes on values $x_1..x_m$ with probabilities $p(x_1)...p(x_m)$.

- Its expected value is defined as $$E(X) = \sum_i x_i P(x_i)$$

- E.g. In a TV game, you will lose -1000 TL with 0.6 probability; and gain 10,000 TL with 0.4 probability. What is the expected reward?

E[reward]=-1,000x0.6 + 10,000x0.1 = -600+1000 = 400 TL

Expected Value and Variance of a Function

The **average value of a function f(x) under a probability distribution p(x)** is called the **expectation** of f(x).

- Weighted average, where weighting is done according to the probabilities of different values of x.

$$\mathbb{E}[f] = \sum_x p(x) f(x) \qquad \mathbb{E}[f] = \int p(x) f(x)\, \mathrm{d}x$$

The variance of a function f(x) provides **a measure of how much f(x) varies around its mean** E[f(x)].

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

Expectations

$$\mathbb{E}[f] = \sum_x p(x) f(x) \qquad \mathbb{E}[f] = \int p(x) f(x)\, \mathrm{d}x$$

Approximate expectation:
(discrete and continuous)

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

In other words, if we have some samples x from a given distribution along with f(x) values, we can calculate the approximate estimate by computing the average of f(x) obtained for these samples obtained from this distribution.

Note that there will be more samples in the data set where p(x) is larger and fewer samples where p(x) is smaller; hence we do not need p(x) anymore!

Now we are going to look at concepts: variance and co-variance , of one or more random variables, using the concept of expectation.

Variance and Covariance

The variance of a random variable x provides **a measure for how much x varies around its mean** E[x].

$$var[x] = \mathbb{E}[(x - \mathbb{E}[x]^2)]$$

Co-variance of two random variables x and y measures the extent to which they vary together.

$$\mathrm{cov}[x, y] = \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right]$$

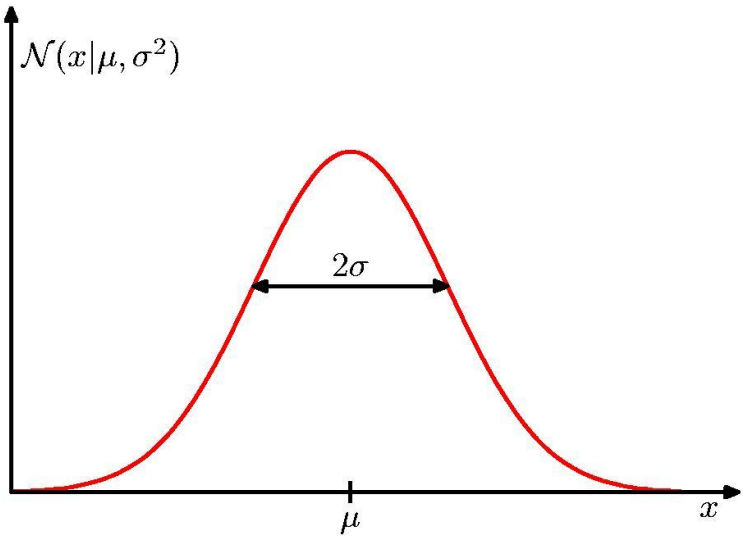
Here the subscript x,y indicates that the expectation is taken over both x and y

Multivariate Normal Distribution

1D Normal (Gaussian) Distribution

1D Normal distribution is defined by its mean μ and variance σ^2 denoted as $N(x|\mu,\sigma^2)$ – it gives the probability density at any point x under Normal distribution with the given two parameters.

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



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

Expectations

Definition of expectation: $\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x$

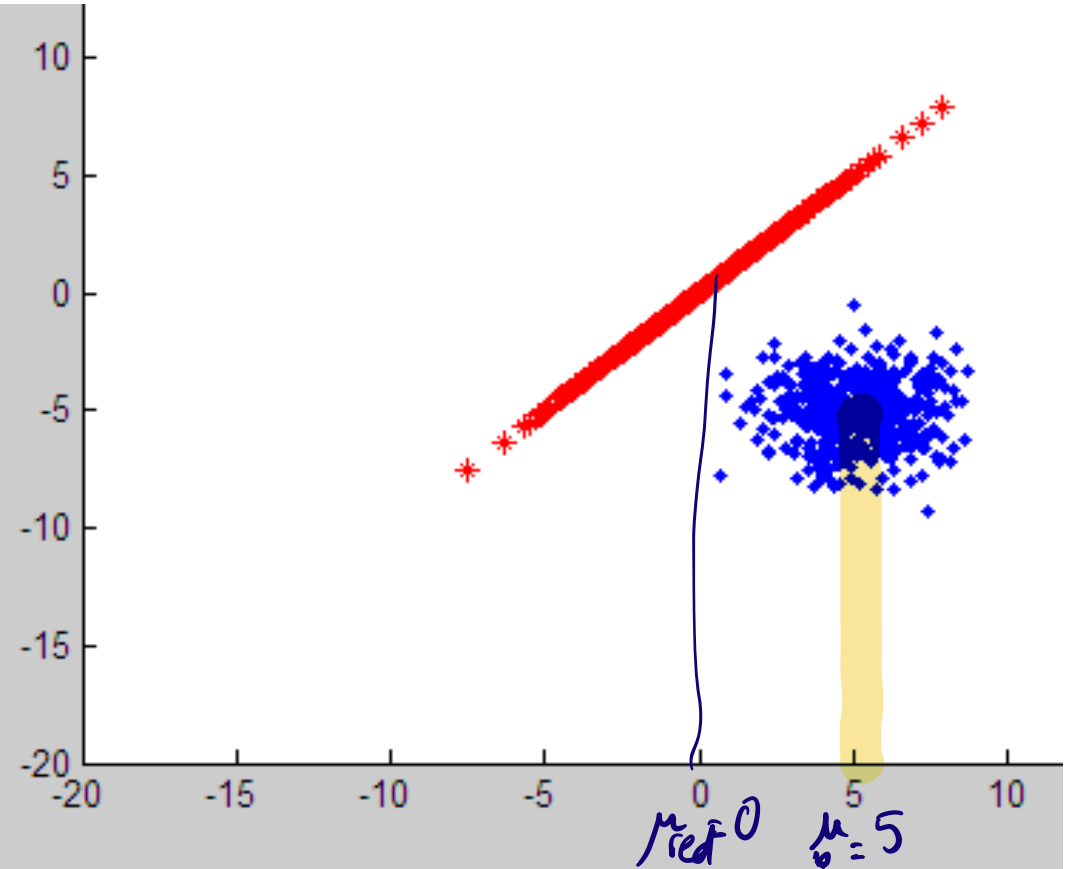
For normally distributed x:

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \, \mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

- We will see how can we characterize p(**x**), assuming **x** is normally distributed.
  - For 1-dimension, we need the mean (μ) and variance (σ^2)
  - For d-dimensions, we need
    - the d-dimensional **mean** vector μ
    - dxd dimensional **covariance** matrix Σ



Multivariate Normal Distribution

- For a single variable, the normal density function is:

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

- For variables in higher dimensions, this generalizes to:

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

where the mean μ is now a d-dimensional vector,

$$\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}]$$

$$\boldsymbol{\Sigma} = \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}].$$

Understanding the Multivariate Normal Distribution

Multivariate Parameters: Mean, Covariance

$$\text{Mean}: E[\mathbf{x}] = \boldsymbol{\mu} = [\mu_1, \ldots, \mu_d]^T$$

$$\Sigma \equiv \text{Cov}(\mathbf{x}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & & & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

$$\text{where } \sigma_{ij} \equiv \text{Cov}(x_i, x_j) \equiv E[(x_i - \mu_i)(x_j - \mu_j)]$$

and σ_i^2 is the variance of the ith feature.

Multivariate Parameters: Mean, Covariance

$$\text{Mean}: E[\mathbf{x}] = \boldsymbol{\mu} = [\mu_1, \ldots, \mu_d]^T$$

$$\Sigma \equiv \text{Cov}(\mathbf{x}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & & & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

In matrix notation:

$$= E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = E[\mathbf{x}\mathbf{x}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

where $\sigma_{ij} \equiv \text{Cov}(x_i, x_j) \equiv E[(x_i - \mu_i)(x_j - \mu_j)]$

$$\sigma_{ij} = \frac{1}{N} \sum_{n=1}^{N} (x_i^{(n)} - \mu_i)(x_j^{(n)} - \mu_j)$$

Matlab code

```
close all;
rand('twister', 1987)       % seed

%Define the parameters of two  2d-Normal distribution
mu1 = [5 -5];
mu2 = [0 0];
sigma1 = [2 0; 0 2];
sigma2 = [5 5; 5 5];

N=500; %Number of samples we want to generate from this distribution

samp1 = mvnrnd(mu1,sigma1, N);
samp2 = mvnrnd(mu2, sigma2, N);

figure; clf;
plot(samp1(:,1), samp1(:,2),'.', 'MarkerEdgeColor', 'b');
hold on;
plot(samp2(:,1), samp2(:,2),'*', 'MarkerEdgeColor', 'r');
axis([-20 20 -20 20]); legend('d1', 'd2');
```

blue

```
mu1 = [5 -5];
mu2 = [0  0];
```

red

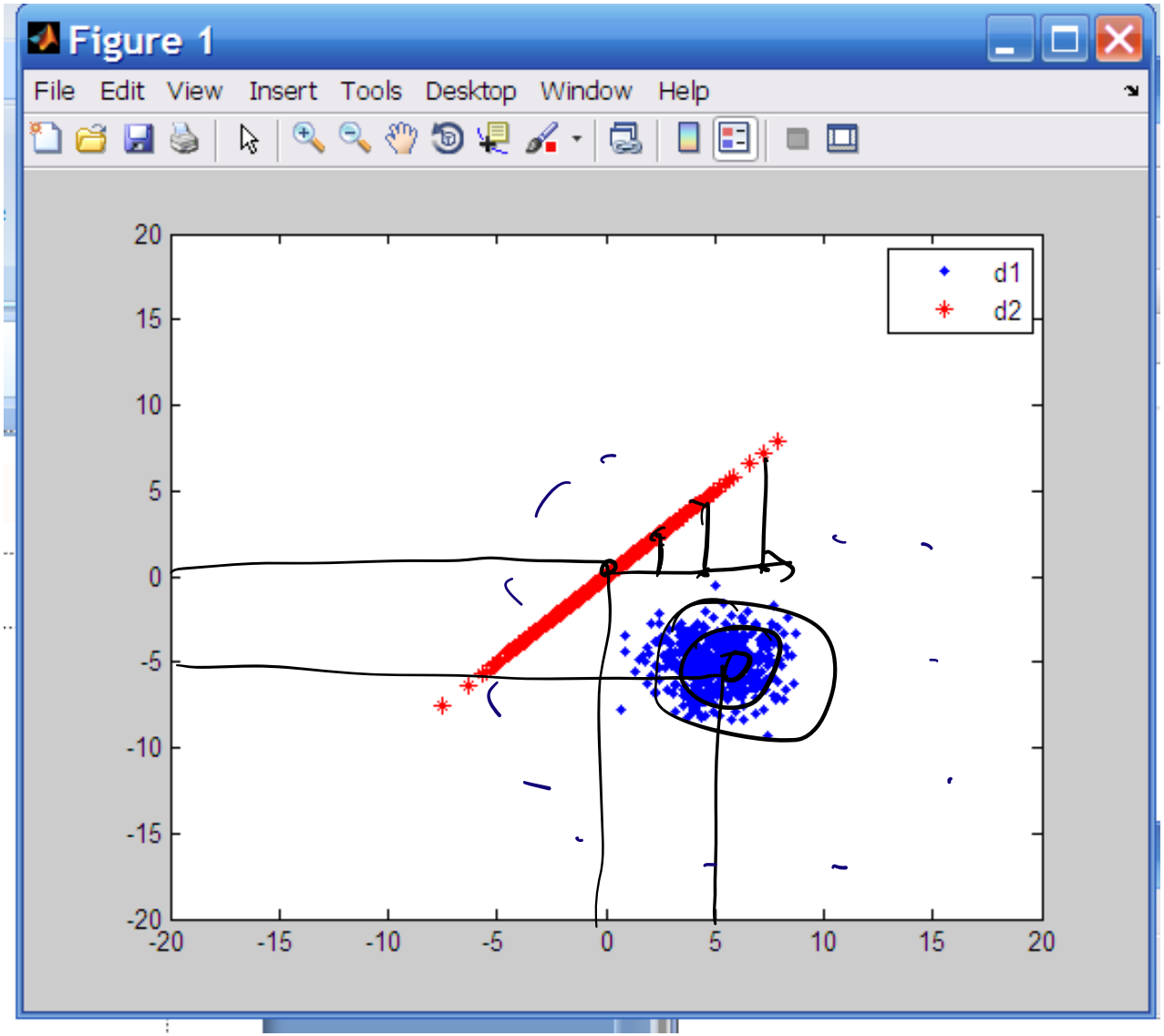
```
sigma1 = [2 0; 0 2];
sigma2 = [5 5; 5 5];
```

$$\Sigma_{blue} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\Sigma_{red} = \begin{bmatrix} 5 & 5 \\ 5 & 5 \end{bmatrix}$$

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \cdot \sigma_2}$$

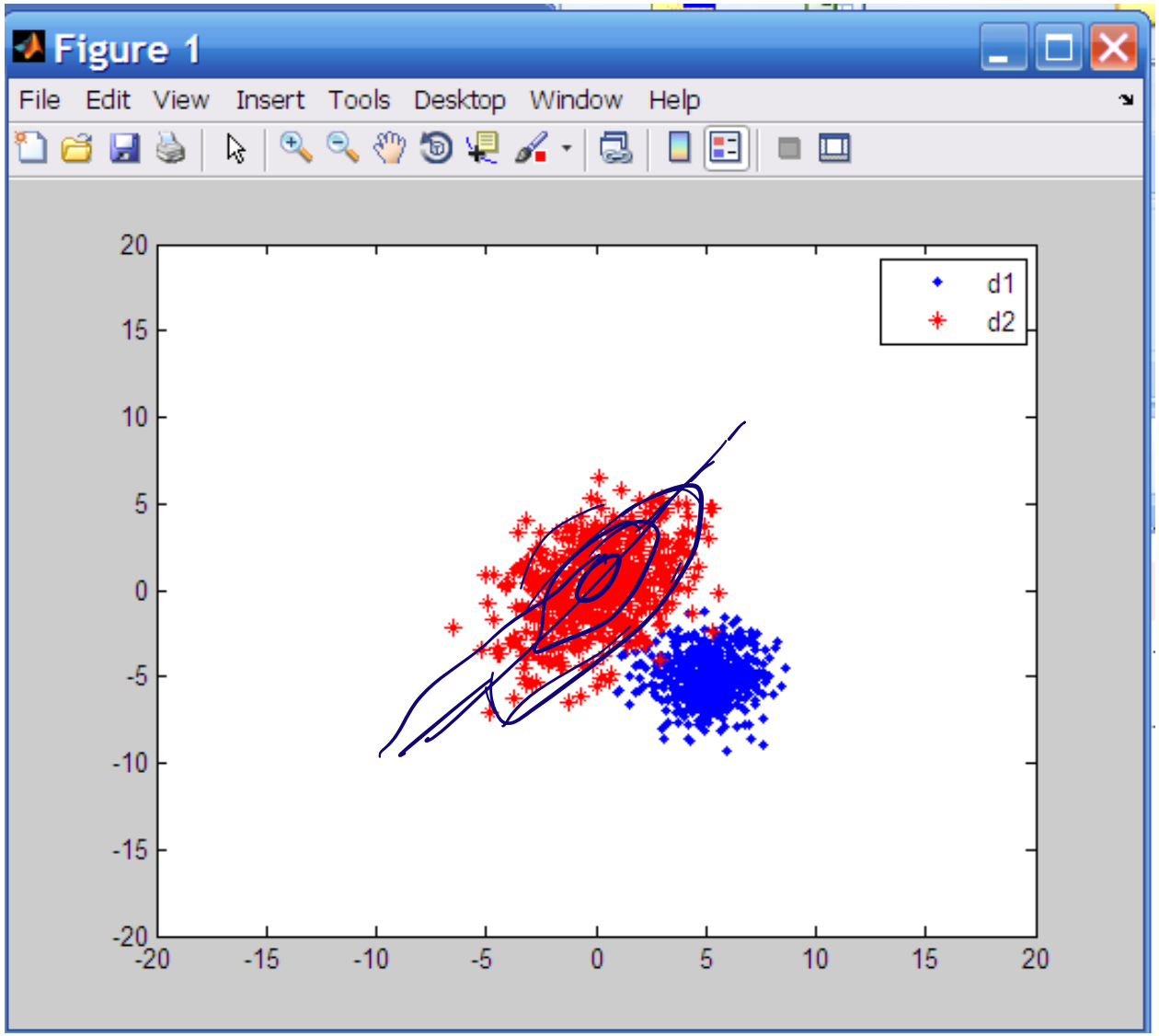
correlation



```
mu1 = [5 -5];
mu2 = [0  0];

sigma1 = [2 0; 0 2];
sigma2 = [5 2; 2 5];
```

$$\Sigma_{red} = \begin{bmatrix} 5 & 2 \\ 2 & 5 \end{bmatrix}$$



Matlab sample cont.

- % Lets compute the mean and covariance as if we are given this data
- **sampmu1 = sum(samp1)/N;**
- **sampmu2 = sum(samp2)/N;**
- sampcov1 = zeros(2,2);
- sampcov2 = zeros(2,2);
- **for i =1:N**
- **sampcov1 = sampcov1 + (samp1(i,:)-sampmu1)' * (samp1(i,:)-sampmu1);**
- **sampcov2 = sampcov2 + (samp2(i,:)-sampmu2)' * (samp2(i,:)-sampmu2);**
- **End**
- sampcov1 = sampcov1 /N;
- sampcov2 = sampcov2 /N;
- %%
- % Lets compute the mean and covariance as if we are given this data USING MATRIX OPERATIONS
- % Notice that in samp1, samples are given in ROWS – but for this multiplication, columns * rows is req.
- **sampcov1 = (samp1'*samp1)/N - sampmu1'*sampmu1;**
- %Or simply
- **mu=mean(samp1);**
- **cov=cov(samp1);**

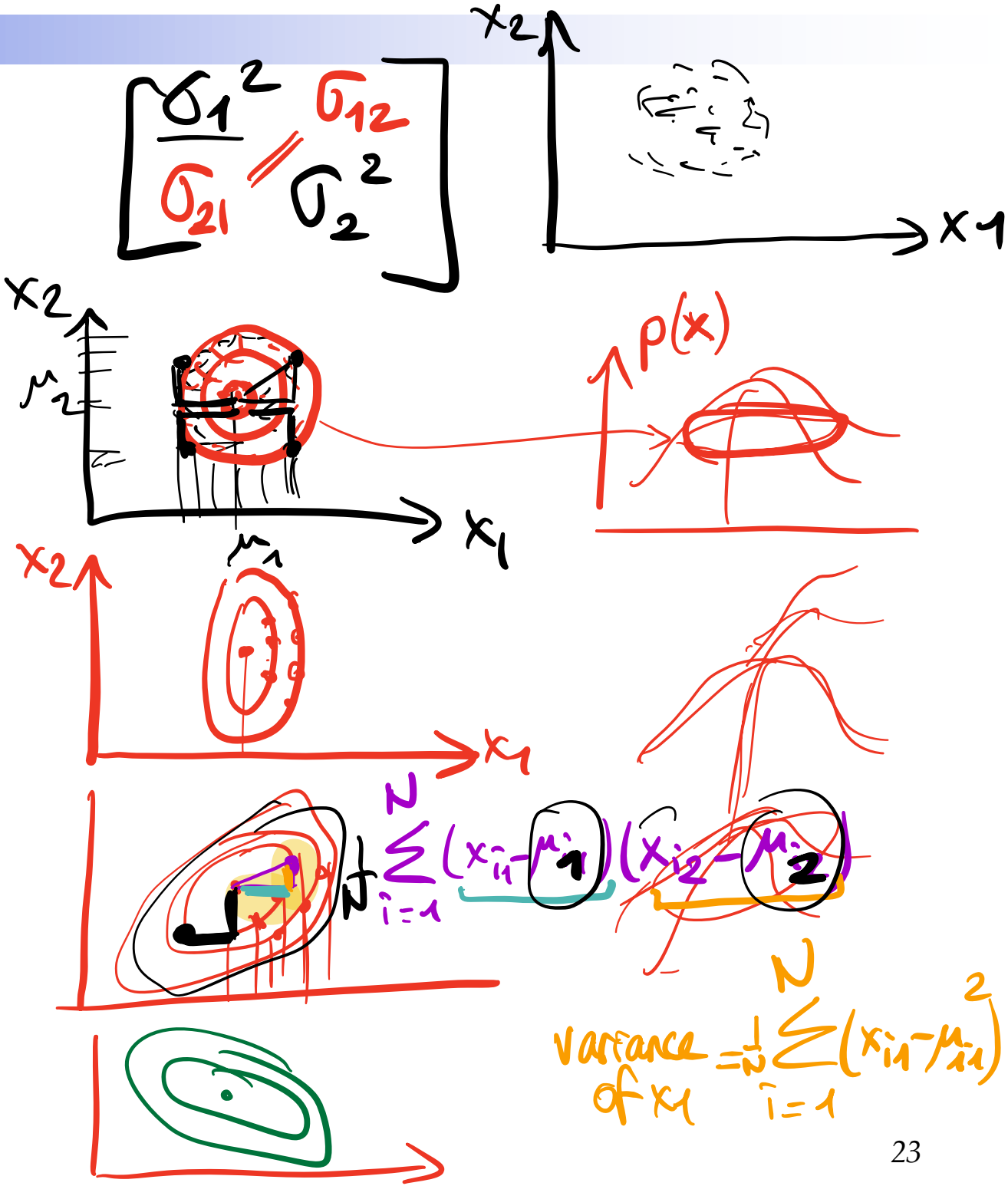
- Some cases:

- A) $\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$

- B) $\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 9 \end{bmatrix}$

C) $\Sigma = \begin{bmatrix} 4 & 2 \\ 2 & 9 \end{bmatrix}$

D) $\Sigma = \begin{bmatrix} 4 & -2 \\ -2 & 9 \end{bmatrix}$



- **Variance**: How much X varies around its expected value

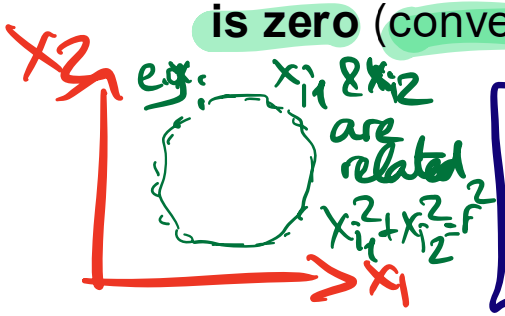
$$\text{Covariance}: \sigma_{ij} \equiv \text{Cov}(x_i, x_j) \equiv E[(x_i - \mu_i)(x_j - \mu_j)]$$

$$\text{Correlation}: \text{Corr}(x_i, x_j) \equiv \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

- **Covariance** is the measure the strength of the **linear relationship** between two random variables
  - covariance becomes more positive for each pair of values which differ from their mean in the same direction
  - covariance becomes more negative with each pair of values which differ from their mean in opposite directions.
  - **if two variables are independent, then their covariance/correlation is zero** (converse is not true).

- **Correlation** is a dimensionless measure of linear dependence.
  - **range between –1 and +1**

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \cdot \sigma_j} = \frac{5}{\sqrt{5} \cdot \sqrt{5}} = +1.$$



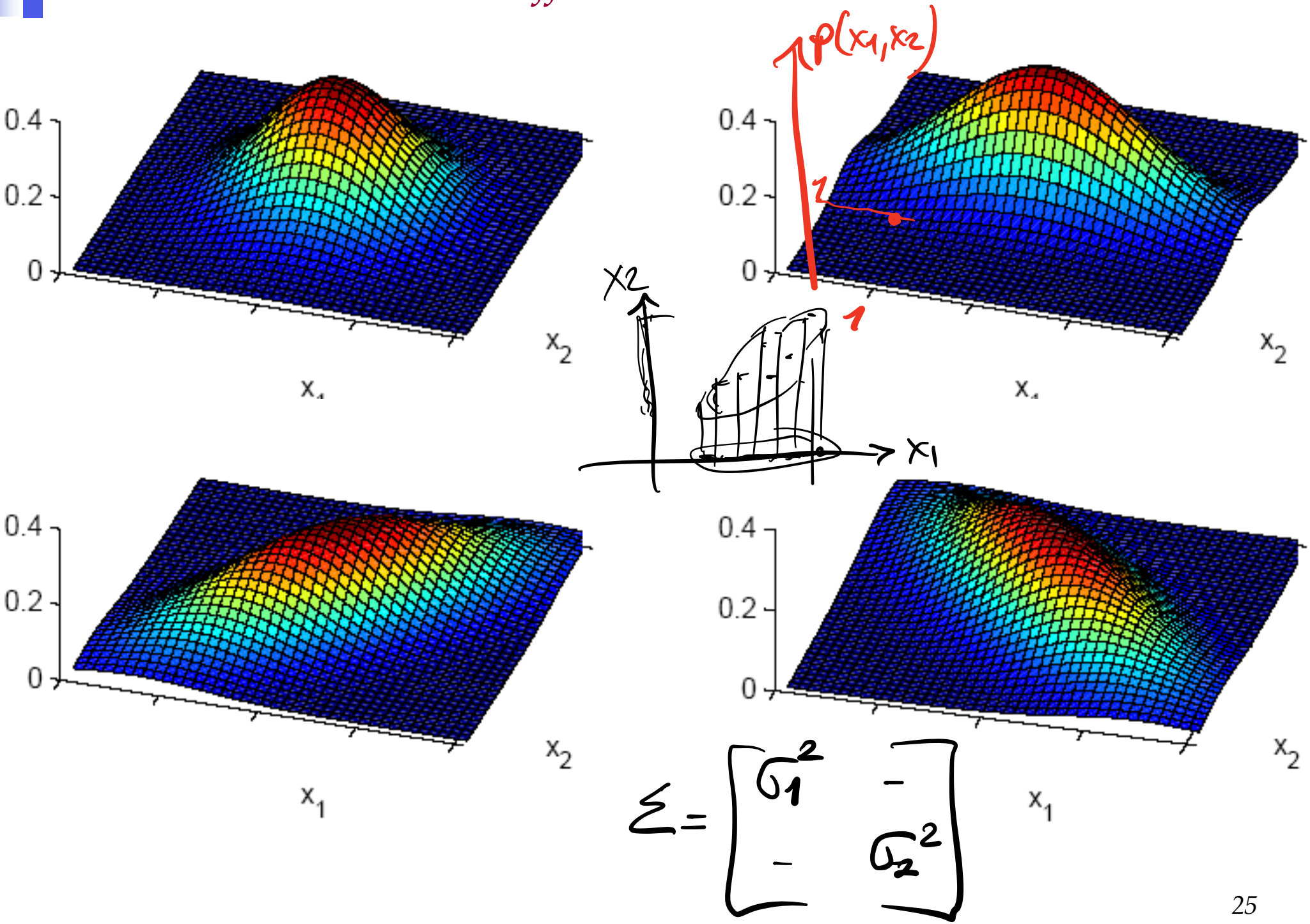
$$\begin{bmatrix} 5 & -5 \\ -5 & 5 \end{bmatrix} \rightarrow \rho_{12} = -1$$

$$\Sigma_{bad} = \begin{bmatrix} 5 & 5 \\ 5 & 5 \end{bmatrix}$$

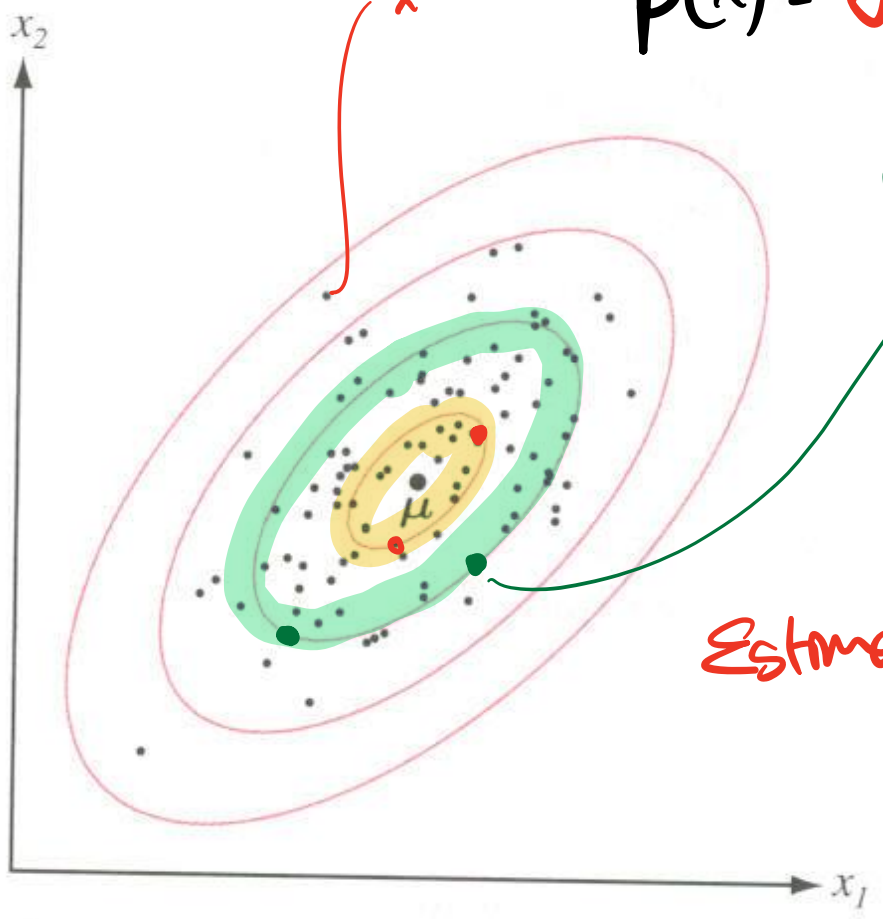
$\sigma_2^2 = 5$

$\Rightarrow \sigma_2 = \sqrt{5}$

How to characterize differences between these distributions



Shape and orientation of the hyper-ellipsoid centered at μ is defined by Σ



$p(x) = \mathcal{N}(x|\mu, \Sigma)$

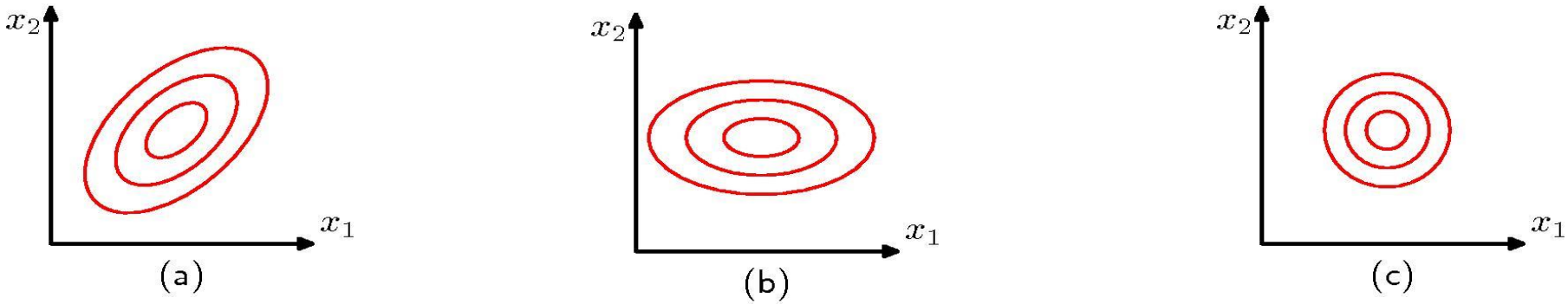
x

$p(x_{\text{green contour}}) < p(x_{\text{red contour}})$

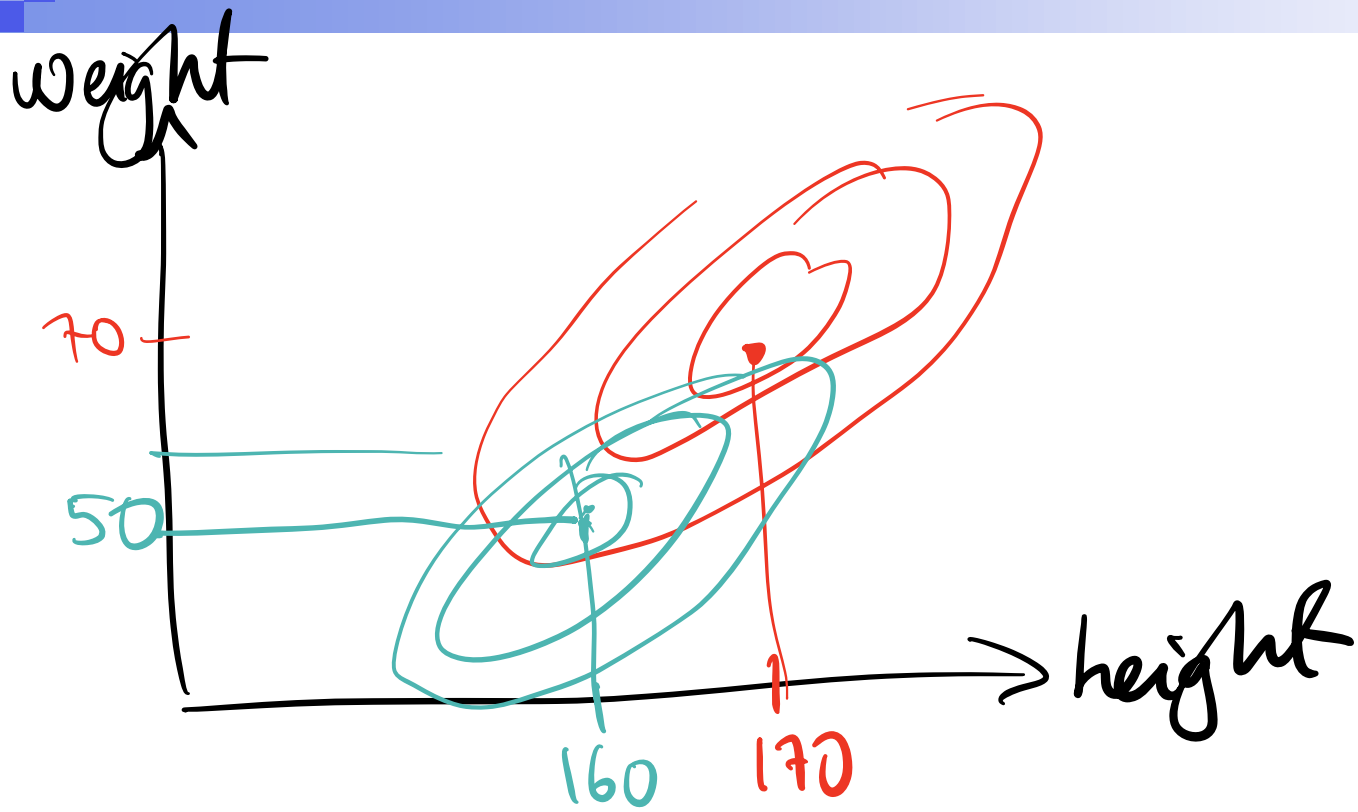
$\mu = \begin{bmatrix} \mu_1 \\ \mu_d \end{bmatrix}$ Estimate

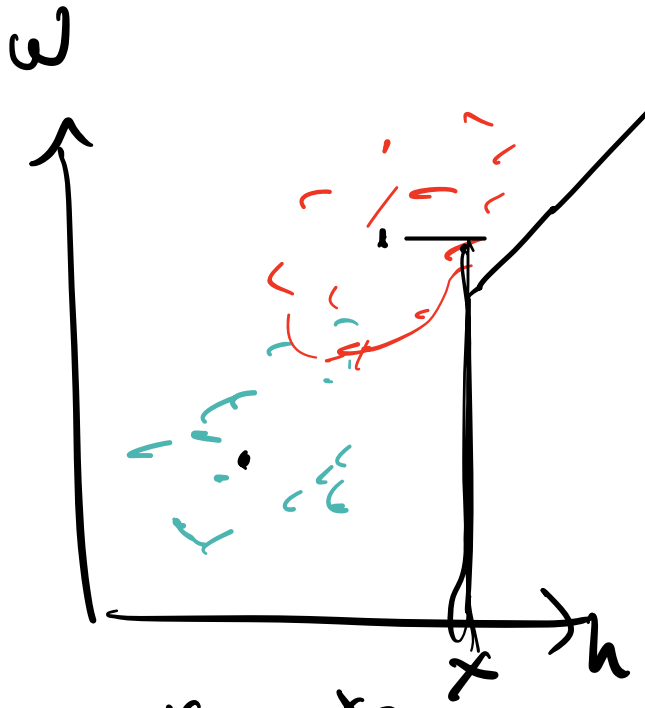
$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1d} \\ & & \sigma_d^2 \end{bmatrix}$

FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The ellipses show lines of equal probability density of the Gaussian.



Contours of constant probability density for a 2D Gaussian distributions with
a) general covariance matrix
b) diagonal covariance matrix (covariance of x1,x2 is 0)
c) Σ proportional to the identity matrix (covariances are 0, variances of each dimension are the same)





$$x = \begin{matrix} x_1 & x_2 \\ [180 & 70] \end{matrix}$$

Mahalanobis Distance

$P(\text{male} \mid x)$?

$P(\text{female} \mid x)$

$P(\text{male})$

$P(x \mid \text{male})$

estimate $\mu_{\text{male}}, \Sigma_{\text{male}}$ to estimate $p(x \mid \text{male})$

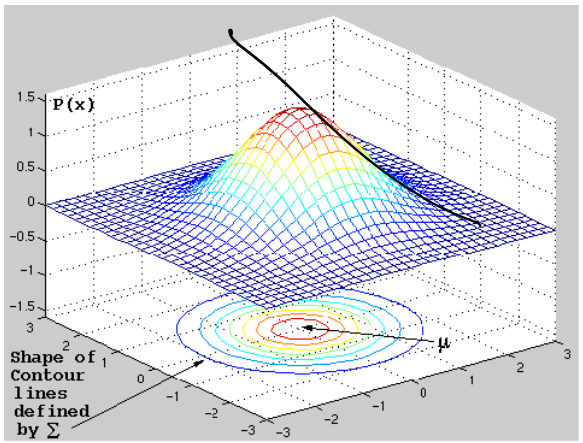
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

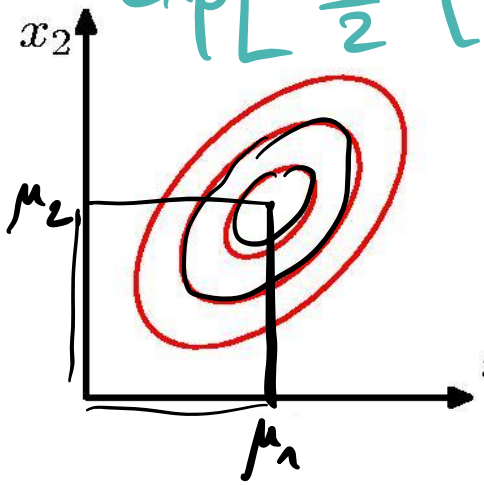
: multivariate

one sample

Mahalanobis Distance

$\exp\left[\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$: 1D





$[x-\mu^T][\Sigma^{-1}][x-\mu]$ = scalar

e.g $[x_1-\mu_1 \cdots x_d-\mu_d]$

From the equation for the normal density, it is apparent that points which have the **same density** must have the same constant term:

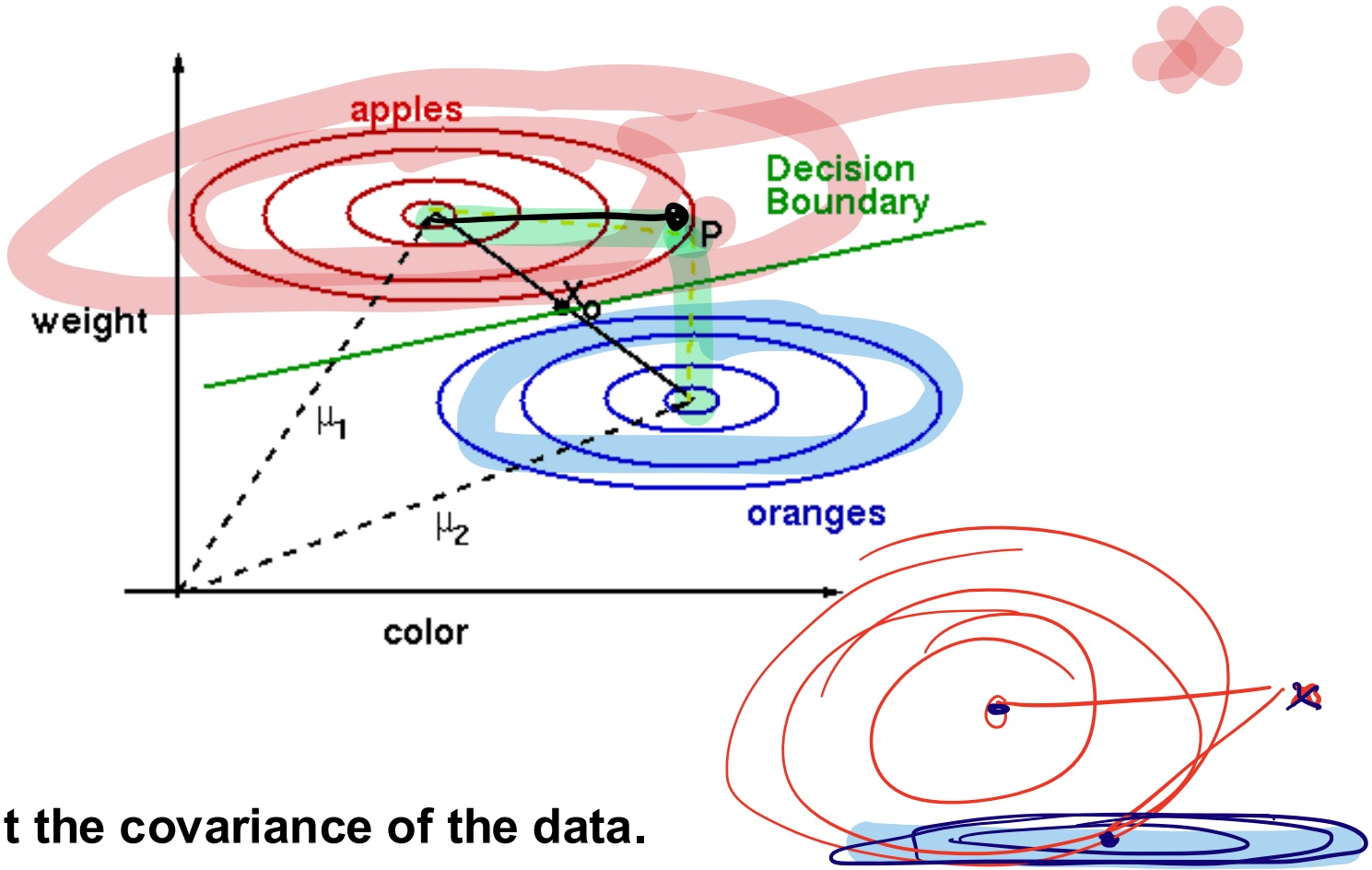
$(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ = 180

$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix}$

e.g $x = \begin{bmatrix} x_1 = 180 \\ \vdots \\ x_d = 70 \end{bmatrix}$ $x-\mu = \begin{bmatrix} x_1-\mu_1 \\ \vdots \\ x_d-\mu_d \end{bmatrix}$

Mahalanobis distance measures the distance from x to μ in terms of ∑

Why Mahalanobis Distance



It takes into account the covariance of the data.

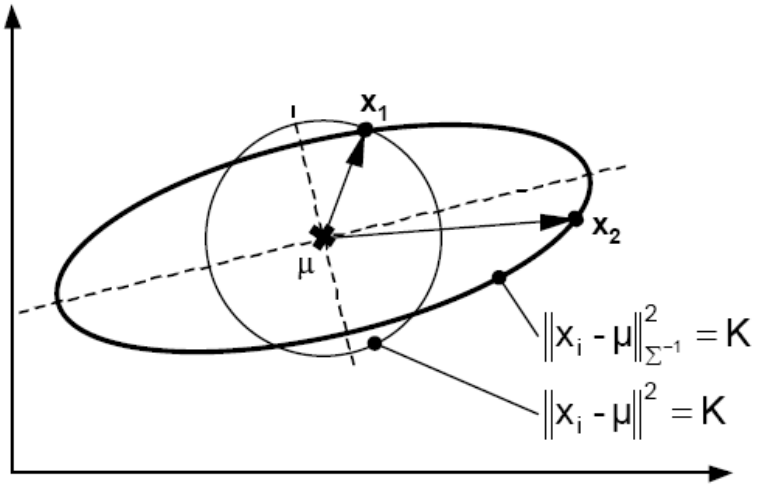
- Point P is at closer (Euclidean) to the mean for the orange class, but using the Mahalanobis distance, it is found to be closer to 'apple' class.

Points that are the same distance from μ

- The quadratic term is called the Mahalanobis distance, a very important distance in Statistical PR

Mahalanobis Distance

$$\|x - y\|^2_{\Sigma^{-1}} = (x-y)^T \Sigma^{-1} (x-y)$$



- **The Mahalanobis distance is a vector distance that uses a Σ^{-1} norm**
  - Σ^{-1} can be thought of as a stretching factor on the space
  - Note that for an identity covariance matrix (Σ=I), the Mahalanobis distance becomes the familiar Euclidean distance

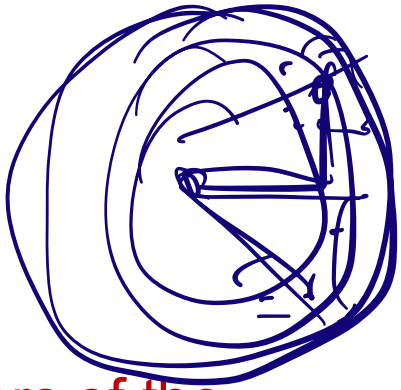
•The ellipse consists of points that are equi-distant to the center w.r.t. Mahalanobis distance.

•The circle consists of points that are equi-distant to the center w.r.t. The Euclidian distance.

Overview

- **Gaussian Bayes classifier** …
  - is a **Bayesian classifier** so it assigns the input **x** to the class C_j for which $P(C_j|\mathbf{x})$ is largest;
  - assumes that **x** is **multivariate normal.**

    $p(\mathbf{x} \mid C_j) = N(\mathbf{x} \mid \mu_j, \Sigma_j)$



  - **It obtains $p(\mathbf{x} \mid C_j)$** by estimating the parameters of the multivariate Normal distribution from the data.

    μ_j and Σ_j

    - typically using the Maximum Likelihood Estimation
  - **Uses Bayes rule for decision**

$$C_i = \arg\max_{C_i} p(x|C_i) \times P(C_i)$$

Example: Gaussian case, μ unknown

- **Assume a dataset $X=\{x^{(1}, x^{(2},\ldots x^{(N}\}$ and a density of the form $p(x)=N(\mu,\sigma)$ where the standard deviation σ is known**
- **What is the Maximum Likelihood estimate of the mean?**

$$\theta=\mu \Rightarrow \hat{\theta}=\operatorname{argmax}\sum_{k=1}^{N}\log p\left(x^{(k}\mid\theta\right)$$

$$=\operatorname{argmax}\sum_{k=1}^{N}\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}\left(x^{(k}-\mu\right)^2\right)\right)$$

$$=\operatorname{argmax}\sum_{k=1}^{N}\left\{\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)-\frac{1}{2\sigma^2}\left(x^{(k}-\mu\right)^2\right\}$$

- The maxima (or minima) of a function are defined by the zeros of its derivative:

$$\frac{\partial\sum_{k=1}^{N}\log p\left(x^{(k}\mid\theta\right)}{\partial\theta}=\frac{\partial}{\partial\mu}\sum_{k=1}^{N}\{\bullet\}=0\Rightarrow\boxed{\mu=\frac{1}{n}\sum_{k=1}^{N}x^{(k}}$$

- So the ML estimate of the mean is the average value of the training data, a very intuitive result!

Maximum (Log) Likelihood for a 1D-Gaussian

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

We find that the estimates found by ML estimation, namely μ_{ML} and σ_{ML} are the sample mean and sample variance.
The actual μ and σ are called population mean and variance.

When you tell scikit to generate samples from a 1D normal distribution with μ and σ, you give population params; when you estimate them from those samples usng the maximum likelihood method, you obtain the sample mean and var – which coincide with the ML estimates for the population params.

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2$$

Maximum Likelihood Estimation for **Multivariate Normal Distribution**

Using the notation in Ethem book:

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N}$$

$$\boldsymbol{m}_i = \frac{\sum_t r_i^t \boldsymbol{x}^t}{\sum_t r_i^t}$$

$$\mathbf{S}_i = \frac{\sum_t r_i^t \left(\boldsymbol{x}^t - \boldsymbol{m}_i\right)\left(\boldsymbol{x}^t - \boldsymbol{m}_i\right)^T}{\sum_t r_i^t}$$

where r_i^t is 1 if the t_{th} sample belongs to class i

Using r_i^t is equivalent to saying:
Take into consideration the sample x^t only if it belongs to class i

Summary

- Naïve Bayes Classifier with **Categorical Attributes**
  - Estimate $P(x_j \mid C_i)$ using frequencies and Laplace smoothing
- Naïve Bayes Classifier with *a* **Single Continuous** Attribute x_1
  - Estimate μ and σ for the single attribute
  - Use $p(x_1 \mid C_i) = N(x_1 \mid \mu,\sigma)$ in the Bayes formula
- Naïve Bayes Classifier with **Multiple Continuous** Attributes
  - Estimate $\boldsymbol{\mu_i}$ (d-dimensional) **and** $\boldsymbol{\Sigma_i}$ (dxd dimensional) for all classes C_i
  - Use $p(\mathbf{x} \mid C_i) = N(\mathbf{x} \mid \boldsymbol{\mu_i},\boldsymbol{\Sigma_i})$
- Further simplifications tomorrow…