# Phan Viet Hoang

# Cinnamon AI Bootcamp 2020

# Assignment 1

## Some useful equations

In general, the Multivariate Gaussian density function is given by:

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

Where x represents our data points,
D is the number of dimensions of each data point,
μ and Σ are the mean and covariance, respectively

**!! Notice** that I'm providing the proof for the case of n-dimension data, so the notation has changed

For later purposes, we will also find it useful to take the log of this equation, which is given by:

$$\ln \mathcal{N}(\mathbf{x}|\mu,\Sigma) = -\frac{D}{2}\ln 2\pi - \frac{1}{2}\ln \Sigma - \frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)$$

In a Bayesian approach, parameters **θ = (π, μ, Σ)** are assumed to be random

Bayes’s rule shows that

$$\begin{aligned} P(\theta|\text{Data}) &\equiv P(\theta|X) \\ &\propto P(\theta)\, P(X|\theta) \\ \text{posterior} &\propto \text{prior} \times \text{likelihood} \qquad \textbf{(*)} \end{aligned}$$

## Back to our problem

### 3/ Please prove equation (2) using Bayes’ theorem

From the product rule of probabilities, we know that

$$p(\mathbf{x}_n, \mathbf{z}) = p(\mathbf{x}_n|\mathbf{z})p(\mathbf{z})$$

sum up the terms on z, we obtained:

$$p(\mathbf{x}_n) = \sum_{k=1}^{K} p(\mathbf{x}_n|\mathbf{z})p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)$$

likelihood function:

$$p(\mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{x}_n) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)$$

Hence

$$\ln p(\mathbf{X}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)$$

Follow Bayes theorem, we have:

$$p(z_k = 1|\mathbf{x}_n) = \frac{p(\mathbf{x}_n|z_k = 1)p(z_k = 1)}{\sum_{j=1}^{K} p(\mathbf{x}_n|z_j = 1)p(z_j = 1)}$$

We also have :

$$p(z_k = 1) = \pi_k, \qquad p(\mathbf{x}_n|z_k = 1) = \mathcal{N}(\mathbf{x}_n|\mu_k, \mathbf{\Sigma}_k)$$

Therefore:

$$p(z_k = 1|\mathbf{x}_n) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j)} = \gamma(z_{nk})$$

That is what we want to prove in (2) formula in the case of n-D data

Also derived from the (*) equation we have:

Prior probability: $\pi_k$

Posterior probability: $\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j)}$ :

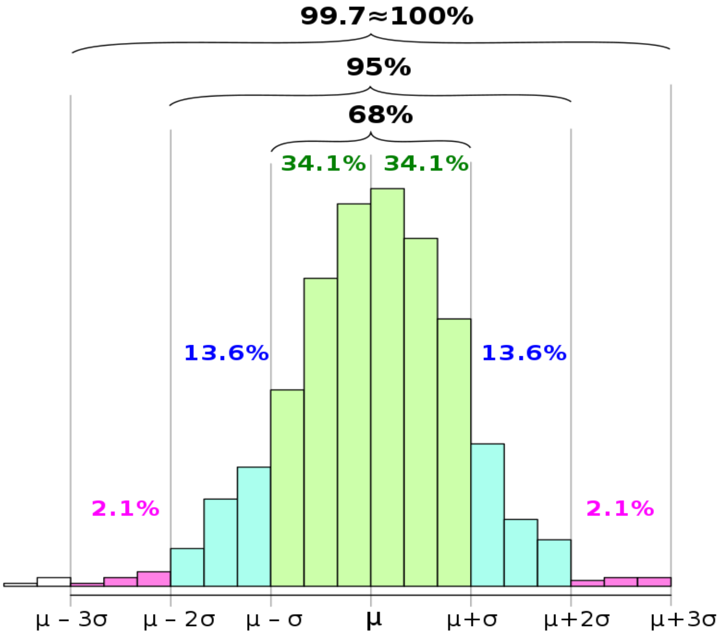
## 4/ Which K do you think is suitable for this problem? Can you come up with a threshold formula based on the normal samples? Justify your answer.

- With a given imbalanced data set of normal and abnormal samples, we will assume the observations in the main class as normal data, and use only these data to train our model. Then, we are able to predict whether a new observation is normal. Therefore K=1 is the most suitable value

- GMM provide us with the probabilities that a given data point belongs to each of the possible clusters.

- The cut-off threshold chosen in pathologic detection problems is relatively high in order to avoid abnormal samples missing(i.e archives high recall) - which is really necessary. That is the reason why I chose the threshold equal to the mean probability taken over all the normal samples

( Others possible cut-off thresholds can be chosen by the 68–95–99.7 rule to modify the model but as we discussed before, there will be a trade-off between TPR and TNR)


99.7≈100%
95%
68%
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
μ - 3σ
μ - 2σ
μ - σ
μ
μ+σ
μ+2σ
μ+3σ