# ECE 0402 - Pattern Recognition

Lecture 4 on 1/26/2022

**Review:**

Consider $(X, Y)$ pair where

- $X$ is a random vector in $\mathbb{R}^d$
- $Y \in \{0, 1, ..., K-1\}$ is a random variable that depends on $X$.

Let $f : \mathbb{R}^d \to \{0, 1, ..., K-1\}$ be a some classifier with **probability of error/risk** given by

$$R(f) \;:=\; \mathbb{P}[f(X) \neq Y]$$

The **Bayes classifier**, $f^*$, is the optimal classifier – with smallest risk possible.

We can calculate the Bayes classifier explicitly if we actually know the joint distribution of $(X, Y)$. In general we can't do this but it is a useful place to start.

- $g(x, k)$ denote the joint distribution of $(X, Y)$.
- Instead we can write this as $g(x, k) = \pi_k \; g_k(x)$
  - $\pi_k \;=\; \mathbb{P}[Y = k]$ is the **a priori probability** of class $k$.
  - $g_k(x)$ is the conditional pdf of $X|Y = k$ , i.e., class conditional pdf.
- In turn, from Bayes' rule we have

marginal of x

$\pi_\ell$ : $P[Y = \ell]$

$$\pi_k g_k(x) = \;\pi_k\Big(\frac{\mathbb{P}[Y = k|X = x]\;(\sum_{l=0}^{K-1} \pi_l g_l(x))}{\pi_k}\Big)$$

Maximize $\eta_k(x)$ → minimize risk

$$= \eta_k(x)\;(\sum_{l=0}^{K-1} \pi_l g_l(x))$$

Marginal distribution of x

- $\eta_k(x)$ is the **a a posteriori probability of class** $k$
- $\sum_{l=0}^{K-1} \pi_l g_l(x)$ is just the marginal pdf of x.

maximizing over class k

We showed last time that this Bayes classifier $f^*(x) := \arg\max_k \eta_k(x)$ is the optimum thing to do. In other words, $R(f^*) = \underline{min\ R}(f)$ where the minimum is over all possible classifiers.

To calculate the Bayes classifier.Bayes risk, we need to know $\eta_k(x)$, or equivalently $\pi_k g_k(x)$– which $k$ value makes the $\eta_k$ value biggest or alternatively $\pi_k g_k(x)$.

**Proof:** Consider an arbitrary classifier $f$ and denote the decision regions

$$\Gamma_k(f) := \{x : f(x) = k\}$$

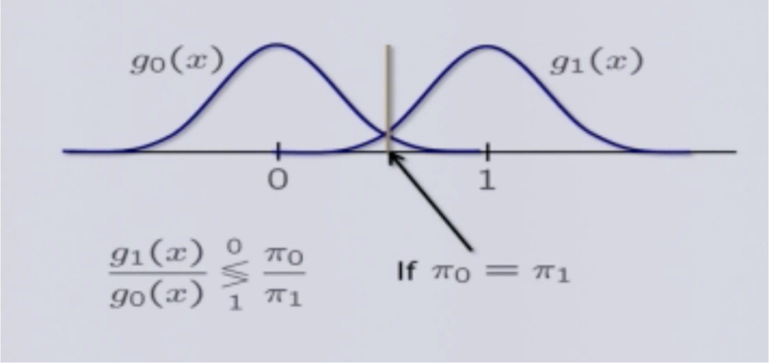
In English, for every $x$ the classifier gives us a label. This $\Gamma_k(f)$ sets are one way to define this classifier. $f$ is breaking up the space of $\mathbb{R}^d$ into these subsets. The reason why this is useful is that this is exactly what you would kind of need to know if you calculate the probability of error. And we said rather than calculating probability of error we calculate:

$$\begin{aligned} 1 - R(f) &= \mathbb{P}[f(X) = Y] \\ &= \sum_{k=0}^{K-1} \pi_k \; \mathbb{P}[f(X) = k | Y = k] \\ &= \sum_{k=0}^{K-1} \pi_k \; \overbrace{\int_{\Gamma_k(f)} g_k(x) dx} \\ &= \int_{\mathbb{R}^d} \sum_{k=0}^{K-1} \pi_k g_k(x) 1_{\Gamma_k(f)}(x) dx \end{aligned}$$

To maximize this, we should select $f$ such that

$$x \in \Gamma_k(f) \;\leftrightarrow\; \pi_k g_k(x) \text{ is maximized}$$

Therefore, the optimal $f$: $f^*(x) := \arg\max_k \underline{\pi_k g_k(x)}$, or equivalently $f^*(x) := \arg\max_k \eta_k(x)$.



which is bigger between $g_0$ and $g_1$ since $\pi_0 = \pi_1$

When could the Bayes risk be zero?

### Generative model

All we have said so far required the knowledge of the joint distribution of $(X, Y)$. But in learning set up, all we know is the training data $(x_1, y_1), ..., (x_n, y_n)$. A generative model is an assumption about the unknown distribution.

Don't have access to full distribution

- typically **parametric**

- build classifier by estimating the parameters via training data
- plug the result into formula for Bayes classifier.

### 1. Linear Discriminant Analysis

We assume

$$X|Y = k \ \sim \mathcal{N}(\mu_k,\ \overset{\downarrow}{\Sigma}) \quad \text{for} \quad k = 0, ..., K-1$$

Assuming each class has Gaussian distribution with different mean and equal covariance

Notice in this assumption the only thing changes between each class is the mean. **Each class has the same covariance matrix $\Sigma$.**

Here $\mathcal{N}(\mu_k,\ \Sigma)$ is multivariate normal distribution with:

$$\phi(x : \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\ |\Sigma|^{1/2}}\ exp\{-\frac{1}{2}(x-\mu)^T\ \Sigma^{-1}\ (x-\mu)\}$$

In LDA, we estimate the prior probabilities $\pi_k$, the mean vectors $\mu_k$, and the covariance matrix $\Sigma$ from the data. Then, we can plug these into our previous formula for the Bayes Classifier (this is why LDA is one of the plug-in methods).

To estimate these from the data, we use

Empirical estimates of:
prior prob.

$$\hat{\pi}_k = \frac{\overset{\text{(Cardinality)}}{|\{i : y_i = k\}|}}{n}$$

mean

$$\hat{\mu}_k = \frac{\sum_{i:y_i=k} x_i}{|\{i : y_i = k\}|}$$

covariance (pool over all classes)

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=0}^{K-1}\sum_{i:y_i=k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T \quad \text{pooled estimate}$$

The LDA classifier is then

$$\begin{aligned}
f(x) &= \arg\max_k \hat{\pi}_k\ \phi(x : \hat{\mu}_k, \hat{\Sigma}) \\
&= \arg\max_k\ log\ \hat{\pi}_k\ +\ log\ \phi(x : \hat{\mu}_k, \hat{\Sigma}) \\
&= \arg\max_k\ log\ \hat{\pi}_k\ -\ \frac{1}{2}(x - \hat{\mu}_k)^T\hat{\Sigma}^{-1}(x - \hat{\mu}_k) \\
&= \arg\min_k\ (x - \hat{\mu}_k)^T\hat{\Sigma}^{-1}(x - \hat{\mu}_k) - 2\ log\ \hat{\pi}_k
\end{aligned}$$

In literature, Mahalanobis distance (between $x$ and $\mu$) is defined as:

$$d_M(x : \mu, \Sigma) = \sqrt{(x - \hat{\mu}_k)^T\hat{\Sigma}^{-1}(x - \hat{\mu}_k)}$$

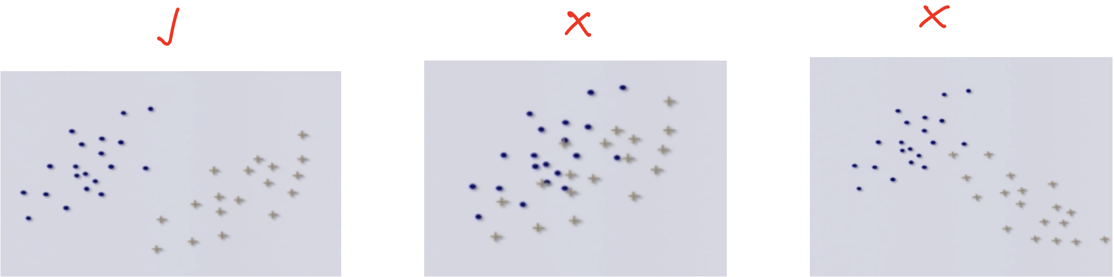


Figure 1: Which sets LDA appropriate?

LDA - assume data comes from Gaussian

Suppose $K = 2$, then

$$d_M(x : \hat{\mu}_0, \hat{\Sigma}) - 2\ log\ \hat{\pi}_0 \ \underset{1}{\overset{0}{\lessgtr}}\ d_M(x : \hat{\mu}_1, \hat{\Sigma}) - -2\ log\ \hat{\pi}_1$$

It turns out that by setting

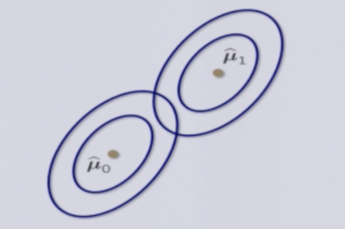
$$\begin{aligned} a &= \hat{\Sigma}^{-1}(\hat{\mu}_0 - \hat{\mu}_1) \\ b &= -\frac{1}{2}\ \hat{\mu}_0^T \hat{\Sigma}^{-1} \hat{\mu}_0 + \frac{1}{2}\ \hat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_1 + log\ \frac{\hat{\pi}_0}{\hat{\pi}_1} \end{aligned}$$

Makes it a linear classifier for k=2

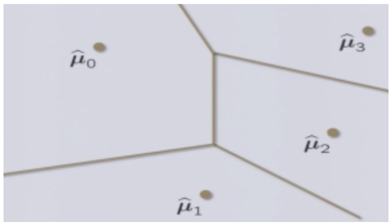
we can re-write the test as:

$$a^T x + b \ \underset{1}{\overset{0}{\lessgtr}}\ 0$$

which in general describes a linear classifier. The contour $\{x : d_M(x : \mu, \Sigma) = c\}$ is an ellipse



What happens if you have more than two classes? The **decision regions** are convex polytopes (intersections of linear half spaces).



Weakness of LDA:

1. The generative model is rarely valid

2. A lot of methods rely on Gaussian distributions are susceptible to outliers.

3. How much data do we need to have idea that this will work? The number of parameters to be estimated is

    - class prior probabilities: $K-1$
    - means: $Kd$
    - covariance matrix: $\frac{1}{2}d(d+1)$

    If $d$ is relatively small and $n$ is large, then we can accurately estimate these parameters (using Hoeffding).

    But $n$ is small an $d$ is large, than we have more unknowns than observations, and will likely obtain poor estimate.

    - as a remedy to this, maybe we can apply dimensionality reduction technique like PCA to reduce $d$
    - or assume a more structured covariance matrix
        - An example of structured covariance matrix: assume $\Sigma = \sigma^2 I$ and estimate $\sigma^2 = \frac{1}{d}tr(\hat{\Sigma})$
        - If $K=2$ and $\hat{\pi_0} = \hat{\pi_1}$, then LDA becomes "nearest centroid classifier":

$$\frac{1}{\hat{\sigma}^2}\|x-\hat{\mu}_0\|^2 \underset{1}{\overset{0}{\lesseqgtr}} \frac{1}{\hat{\sigma}^2}\|x-\hat{\mu}_1\|^2 \implies \|x-\hat{\mu}_0\|^2 \underset{1}{\overset{0}{\lesseqgtr}} \|x-\hat{\mu}_1\|^2$$

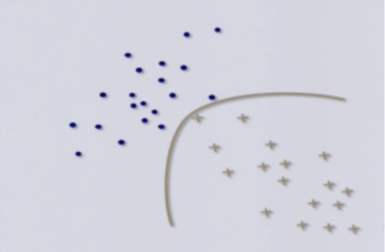
**2. Quadratic Discriminant Analysis**

We may expand the generative model to

$X|Y=k \ \sim \mathcal{N}(\mu_k, \Sigma_k)$ for $k = 0, ..., K-1$

Set $\hat{\Sigma}_k = \frac{1}{|\{i:\ y_i=k\}|} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i-\hat{\mu}_k)^T$

Proceed as before, but this case the decision boundaries will be quadratic.



Are we asking too much with plugin method? Let's have another look:

Suppose $K=2$

Define $\eta(x) = \eta_1(x)$ which is $(= 1 - \eta_0(x))$

In this case, another way to express Bayes classifier is as

$$f^*(x) = \begin{cases} 1 & if\ \eta(x) \geq 1/2 \\ 0 & if\ \eta(x) < 1/2 \end{cases}$$

Note that to express the classifier in this way, we don't actually need to know the full distribution of $(X, Y)$. All we have to know is the distribution of $Y|X$.