

Generative Models for Classification

In the case of [Logistic Regression](#) we directly model $\Pr(Y = k|X = x)$ using the [Sigmoid Function](#) (Logistic Function) its a simple conditional probability approach **Predicting Y given X** .

Considering the alternative Probability of the the the predictors X given a class Y which is the logic behind **Bayes Theorem**, When the distribution of X is normal the model turn to be very similar to Logistic Regression.

Why its needed?

- If There is a substantial separation between the two classes the **parameter estimates** for logistic regression model will be unstable
- If the predictors X are normally distributed and the sample size n is small Using **Generative Models** will be more accurate than Logistic Regression

Suppose that we want to classify an observation into on of the K classes where $K \geq 2$, means the [Response](#) Y can take K possible classes.

- Let π_k be the **Prior Probability** that the a random [Observation](#) comes from the k th class $P(Y = k)$
- Let $f_k(X) = \Pr(X|Y = k)$ is the **Density Function** of X
 - $f_k(x)$ will be large if there is **high probability** of that observation being in the k th class

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum^K \pi_i f_i(x)}$$

- This is just the [Bayes' theorem](#) Formula
- Let $p_k(x) = \Pr(Y = k|X = x)$ and will also be called the **Posterior** probability
- So our goal is to know the probability of the [Observation](#) being in the k th class Given the class k aka [Response](#) Y

Linear Discriminant Analysis Vs Logistic Regression

First before diving into **LDA** ,Lets establish a clear difference between What [Logistic Regression](#) Does and What **LDA Does** , cause at the end they both Classify a new observation into a class.

What Logistic Regression Does

- It tries to directly model the **posterior probability**

$$P(Y = k|X = x) = \frac{e^{\vec{x}\vec{\beta}}}{1 + e^{\vec{x}\vec{\beta}}}$$

- You just want to separate the classes,Its **Discriminative** just making distinctions
- *We don't care how the data X looks like inside each class*(We make no assumption about the distribution of X , We aim class each observation
- We also don't care about how the data was generated

What LDA Does

- Its a **generative**, it models how the data X looks within each class
- With the assumption of the data X taking a normal distribution form

$$P(X|Y = k) = \mathcal{N}(\mu_k, \Sigma)$$

- Then use [Bayes' theorem](#) to compute

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

- Then classify x by choosing the class with the highest **posterior** $P(Y = k|X = x)$
- Simply its saying *This new point looks most like the kind of data that class k generates*

Aspect	Logistic Regression	LDA
Type	Discriminative , Making distinctions	Generative
Models	$P(Y X)$	$P(X Y), P(Y)$
Assumes	No assumptions	$X Y$ is Normally Distributed, All classes variances are equal
Uses	Decision boundaries	Bayes's Classifying

Linear Discriminant Analysis for $p = 1$

For this section we will assume that $p = 1$ only one predictor, The goal is :

- Obtain estimate for $f_k(x) \rightarrow$ The update
- We plug it into the [Bayes' theorem](#) formula to get the estimate for $p_k(x) \rightarrow$ The posterior
- Then we **Classify** an [Observation](#) to the class with the highest $p_k(x)$ The posterior

Estimating $f_k(x)$

Also known as the The update noted $P(X|Y)$ the **easy to measure** probability used to **update** the The prior π_k

- We assume that $f_k(x)$ is **normal or Gaussian**

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k}e^{-1/2\sigma_k^2(x-\mu_k)^2}$$

- $\mu_k \rightarrow$ The mean for the k th class
- $\sigma_k^2 \rightarrow$ the variance for the k th class

Assuming that the variance of all classes K are equal so $\sigma_k^2 = \sigma^2$ for simplicity

$$P(Y = k|X) = p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma}e^{-1/2\sigma^2(x-\mu_k)^2}}{\sum^k \pi_i \frac{1}{\sqrt{2\pi}\sigma}e^{-1/2\sigma^2(x-\mu_k)^2}}$$

- This is the **Posterior** with the assumption of all **variances are equal** and that $f_k(x)$ is **normally distributed**

Dropping the denominator in the **Posterior** formula we get the [Discriminant Function](#) which is the log of the **normalized posterior**, The denominator is constant across all classes k

$$\log(p_k(x)) = \log(\pi_k) - \frac{1}{2\sigma^2}(x - \mu_k)^2$$

- Dropping the term $\frac{1}{\sqrt{2\pi}\sigma}$ for being a constant

$$\log(p_k(x)) = \log(\pi_k) - \frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}\mu_k^2 + \frac{x\mu_k}{\sigma^2}$$

$$\log(p_k(x)) = \log(\pi_k) - \frac{\mu_k}{2\sigma^2} + x\frac{\mu_k}{\sigma^2} = \delta_k$$

- Dropping x^2 will get δ which is the a **Linear Discriminant Function** of x
- δ_k is the score for class k
- We aim to assign $X = x$ to the class that gives the largest δ_k

If $K = 2$ and $\pi_1 = \pi_2$ then the [Bayes decision boundary](#) is the point where $\delta_1 = \delta_2$

$$\delta_1 = \delta_2$$

$$x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \log(\pi_1) = x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \log(\pi_2)$$

$$\log(\pi_1) = \log(\pi_2)$$

$$\begin{aligned} x \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} &= x \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} \\ x \frac{\mu_1 - \mu_2}{\sigma^2} &= \frac{\mu_1^2 - \mu_2^2}{2\sigma^2} \\ x &= \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2} \end{aligned}$$

- If $x > \frac{\mu_1 + \mu_2}{2}$ we assign it to class 1 otherwise class 2

In practice we are not quite certain of our assumption that X is **normally distributed** and we will still need to estimate :

- $\mu_1 \dots \mu_K$
- $\pi_1 \dots \pi_k$
- σ^2