Linear Discriminant Analiisys

Teorema de Bayes para clasificación

- Supongamos que deseamos clasificar una observación en k clases donde k ≥ 2
- puede tomar K posibles valores.

- Sea π_k que representa la probabilidad previa de que una observación elegida al azar sea de la k-esima clase.

- Sea

$$f_k(x) \equiv \Pr(X = x|Y = k)$$

- la función de densidad de X para una observación que viene de la k-esima clase

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \qquad \text{(a)}$$

- Tenemos que estimar π_k y $f_k(x)$, en el caso del primer término básicamente sería la fracción en el training que pertenece a la clase kth. A fin de estimar $f_k(x)$ se plantea algunas suposiciones.

Linear discriminant Analisys para p = 1

- Supongamos que asumimos que $f_k(x)$ es una Normal o Gaussiana:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

- donde μ_k y σ_k^2 son la media y varianza de la k-esima clase y asumimos que $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2$, reemplazando en (a) tenemos,

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)} \qquad \text{(b)}$$

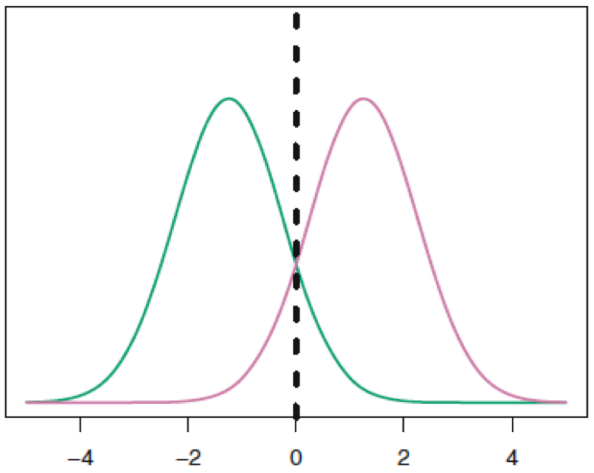
- Así que el clasificador Bayes involucra la asignación de una observación $X = x$ a la clase para la cual es la mayor.

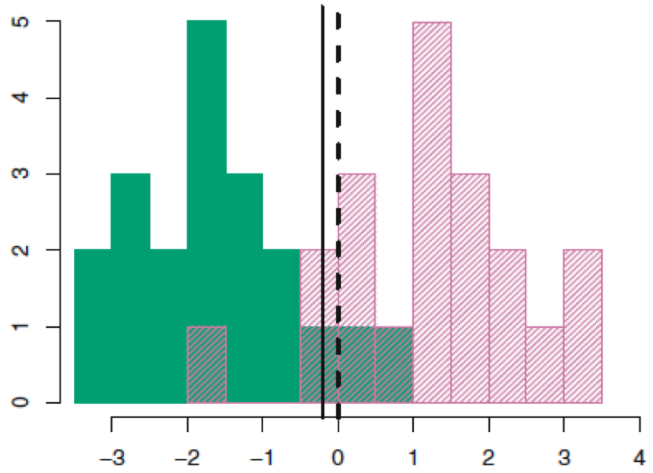
- Aplicando log a (b), tenemos que la siguiente expresión es equivalente a asignar la observación a la clase para la cual

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \qquad \text{(c)}$$

sea la mayor.

- En la siguiente figura las dos funciones de densidad normal $f_1(x)$ y $f_2(x)$ representan dos clases distintas.





- Es así que la técnica de linear discriminant analisys (LDA) aproxima el Clasificador Bayes reemplazando estimadores para π_k, μ_k, y σ^2

- Particularmente se utilizan los siguientes:

$$\begin{aligned} \hat{\mu}_k &= \frac{1}{n_k} \sum_{i:y_i=k} x_i \\ \hat{\sigma}^2 &= \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2 \end{aligned} \qquad \text{(d)}$$

- Con *n* es el número total de observaciones del training set y n_k es el número de observaciones en la i-ésima clase.

- LDA estima π_k usando la proporción de observaciones del training set que pertenece a la *k*th clase, es decir,

$$\hat{\pi}_k = n_k/n. \qquad \text{(e)}$$

- Reemplazando (d), (e) en (c) tenemos,

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k) \qquad \text{(f)}$$

- La palabra linear en el nombre del clasificador viene del hecho de que la función discriminant $\hat{\delta}_k(x)$ es lineal de *x*

- Para implementar el LDA, comenzamos estimado π_k, μ_k, y σ^2 using (d) y (e) que resulta de asignar una observación a la clase para la cual (f) es la mayor.

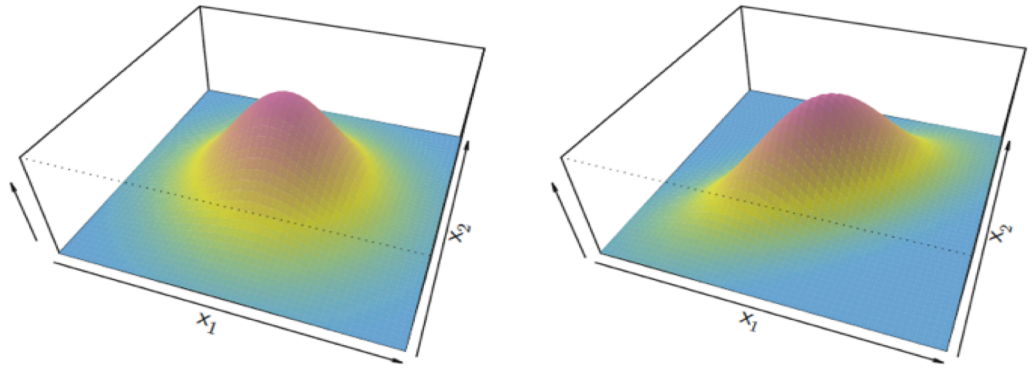
- Entonces el clasificador LDA resulta de asumir que las observaciones dentro de cada clase vienen de una distribución normal con un vector medias y una varianza común σ^2y se reemplaza tales estimadores en el clasificador basado en Bayes.

Linear discriminant Analisys para p > 1

- Ahora, extendiendo el clasificador LDA al caso de multiples predictors. Para lo cual asumimos que X = (X1,X2, . . .,Xp) es obtenida de una distribución Gausiana multivariable con una media de clase específica y una matriz de Covarianza.

- La distribución Gausiana multivariable asume que cada predictor individual sigue una distribución normal de una dimensión.

- Dos ejemplos de una distribución Gausiana multivariable con p = 2 se ,muestra
- en la siguiente figura:



- La figura de la izquierda ilustra el ejemplo en el cual $Var(X_1) = Var(X_2)$ and $Cor(X_1,X_2) = 0$;

- Sin embargo, la forma de campana puede ser distorsionada si los predictores son correlacionados o tengan diferentes varianzas como se observa en. la el lado derecho de la figura arriba.

- Para indicar que una variable aleatoria p-dimensional X tiene una distribución Gausiana multivariable, se denota con $X \sim N(\mu, \Sigma)$. Se tiene $E(X) = \mu$ es la media de X (un vector con p componentes y $Cov(X) = \Sigma$ es la matriz de covarianza p x p de X, formalmente se define como la densidad multivariada Gausiana:

$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right) \qquad \text{(g)}$$

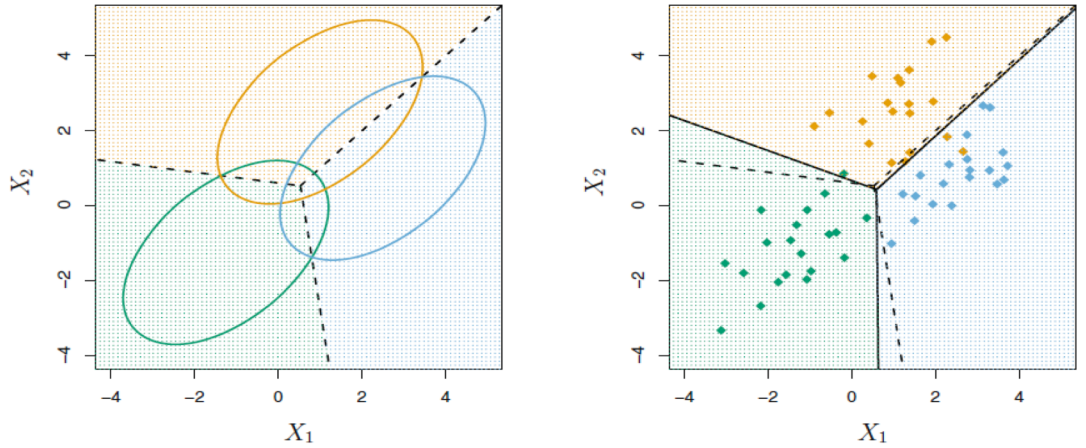
- En el caso de p > 1, el clasificador LDA asume que las observaciones en la *k*-ésima clase son obtenidas de una distribución Gausiana multivariable N(μ, Σ) donde μ_k es un vector específico de medias Σ is a matriz de covarianza que es común a todas las K clases.

- Reemplazando la función de densidad para la kth clase, $f_k(X = x)$, en (a), revela que el clasificador Bayes asigna una observación X = x a la clase para la cual el clasificador Bayes:

$$\delta_k(x) = x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k \qquad \text{(h)}$$

es la mayor. Esta fórmula es una versión Vector/Matriz de (f)

- Un ejemplo se muestra en esta figura de la izquierda. Tres elipses representan regiones que contienen el 95% de probabilidad para cada una de las tres clases.



- Las líneas punteadas corresponden a Bayes Decision Boundaries, en otras palabras ellos representan. el conjunto de valores x fpara lo cual $\hat{\delta}_k(x) = \hat{\delta}_l(x)$; i.e.

$$x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k = x^T\Sigma^{-1}\mu_l - \frac{1}{2}\mu_l^T\Sigma^{-1}\mu_l$$

- para k ≠ l (Notar que $\pi_k$ha desaparecido por que cada una de las tres clases tiene el mismo número de observaciones del training set, i.e. $\pi_k$es el mismo para cada clase.

LDA cuenta con algunas suposiciones a cerca de los datos:

- Asume que los datos tienen una distribución Gausiana
- Cada una de los valores de clase tiene covarianzas idénticas.
- Sin embargo, vale la pena mencionar que inclusive siendo las suposiciones no cumplidas funciona muy bien

Razones por las cuales se requiere Linear Discriminant Analisys teniendo a la Regresión Logística:

- Cuando las clases son bien separadas, la estimación de parámetros para la regresión logística es sorprendentemente inestable, no así LDA.

- Si n es pequeño y la distribución de los predictores es aproximadamente normal en cada una de las clases, LDA es más estable que la regresión logística

- LDA es más popular cuando tenemos más de dos valores de clase.

Bibliografía

- James G., Witten D., Hastie T., Tibshirani R.: “An Introduction to Statistical Learning”, Ed. Springer, 2013.

- Hastie T., Tibshirani R.: “The Elements of Statistical Learning”, Ed. Springer, 2013.