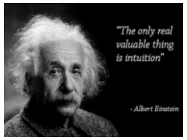
Linear Discriminant Analysis

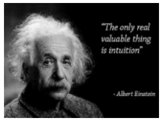


Bayes Theorem: Intuition

- The Bayes Theorem is **fundamental** in statistics. A lot of what we do in predictive modeling is based on **"Bayesian Statistics"**
- The **math** behind Bayes Theorem is beyond the scope of this class, but it helps to understand **intuitively** what it means
- The **Bayes Theorem** describes the **probability** of an **event** based on given **conditions** (that may be related to the event).
- **Non-Bayesian** – e.g., probability (make an A in this class)
  Bayesian – e.g., probability (make an A, given a GPA of 3.5)
- If GPA is **related** to grades, then the **conditional probability** based on **GPA** will be **different** than the **general** probability. In a nutshell ($P(A|B)$ and $P(B|A)$ are conditional probabilities):

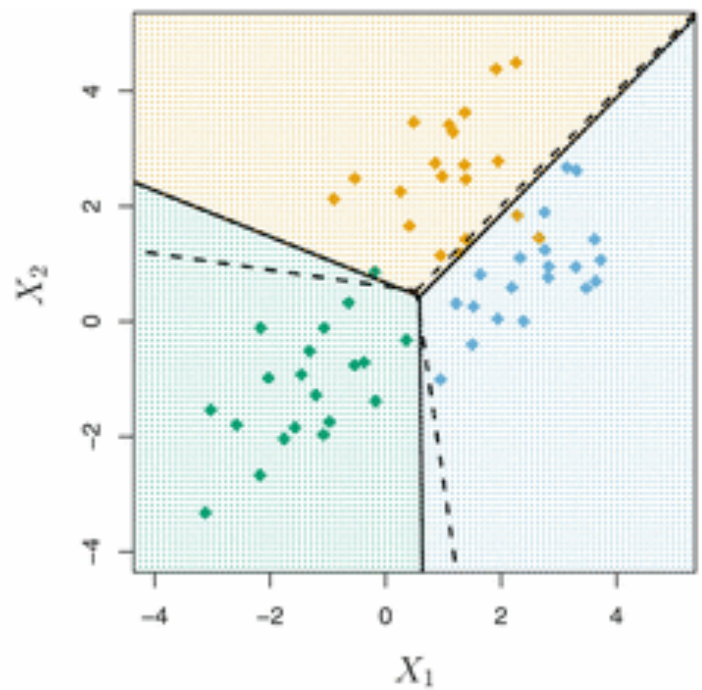
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

E.g., the probability of a student making an A in the class given a GPA>=3.5 is equal to the probability that a student with a GPA>=3.5 making an A, times the probability of anyone making an A, divided by the probability of anyone having a GPA>=3.5



Linear Discriminant Analysis (LDA): Intuition

- Intuitively, LDA is a **simple concept** and it can be best explained with a simple model with an outcome variable Y with $K=3$ classes and $p=2$ predictors X_1 and X_2
- If you plot all the data along the X_1 and X_2 axes and **color-code** the dots for each **category** of Y
- The idea behind **LDA** is to draw **3 lines** (one for each class) and **rotate** the lines until the probability of **correct classification** is maximized
- The lines are called linear **"Discriminant Functions"**
- It is easy to see visually that if the **classes** are easily **separated** LDA works **well**, but **not** so well if classes are **comingled**.



LDA Explained

- If we know the overall **probability** that $Y = k$ (an outcome is in class k) and we know the distribution (i.e., **normal**) of the predictors X we can estimate the probability that $Y = k$ for a given set of X predictors.
- Assuming a **normal distribution** of the X's, we can use **Bayes Theorem** to calculate the probability $P(Y = k \mid X_1 = c_1, X_2 = c_2, etc.)$
- **LDA** estimates a **"discriminant function"** $P(Y = k \mid X_1 = c_1, X_2 = c_2, etc.)$ for all each of the K possible **categories**, assuming that all the X's are **normally** distributed, and **assigns** the observation to the **category** that has the **highest probability**
- The term **LDA** contains the term **"linear"** because the **discriminant function** is a **linear** function of the X's.
- If there are **K** outcome **categories**, there are **K discriminant functions**, one for each K
- **Other types** of discriminant analysis (e.g., quadratic) include **non-linear** functions of the X's.

LDA vs. Logistic Regression

- **LDA** accomplishes the **same** results **as logistic** regression, but through a **probabilistic** method based on **Bayes Theorem**
- **LDA** some times **outperforms logistic** regression in terms of predictive **accuracy**, particularly when:
  1) The outcome is **multinomial** (>2 categorical outcomes) :
  2) The **classification** for each predictor are **well separated**
  3) The **sample is small** and the **distribution** of the individual predictors is approximately **normal**
- There is also a **tradeoff** is with loss of **interpretation**; **logistic regression** is **better** for hypothesis testing and interpretation
- **LDA** estimates a **"discriminant function"** with the **probability** of ***Y*** falling in each of the ***K*** categories for a given set of predictors ***X's***
- **LDA** then **assigns** that observation to the **category** with the **highest probability**
- Alternative modeling should be tested with **cross-validation**

LDA: Fit Statistics

- Like with other classification models, the **confusion matrix** needs to be evaluated to inspect:
  - ➢ **Error Rates** – proportion of incorrect classifications
  - ➢ **Sensitivity** – proportion of correct positive classifications
  - ➢ **Specificity** – proportion of correct negative classifications
- Depending on the **analysis goals**, we may want to place more emphasis on **one** or the **other** – is it better to send the innocent people to jail or to let the guilty go free?
- The Bayes classifier uses the **threshold** probability of **50%** to classify an observation into a category, but this threshold can be **changed** depending on the **analysis goals**.

- Varying this **threshold** and **plotting** the resulting **sensitivity** and **specificity** values yields the **ROC** curve (“Receiver Operating Characteristics”).

KOGOD SCHOOL
of
BUSINESS