Classification - Decision Trees and Bayes Decision Theory

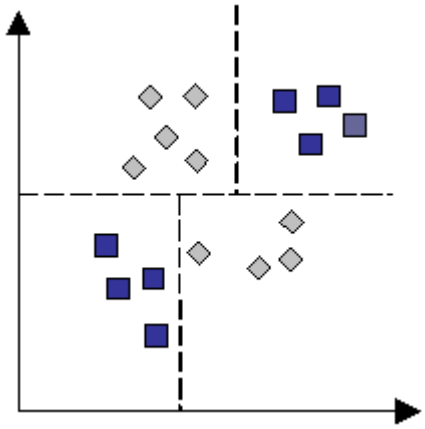
Authors: Justin Wagers (https://www.linkedin.com/in/justin-wagers-47a5a5b3/), Jocelyne Walker (https://www.linkedin.com/in/jocelynewalker/), Rongzhi Xu (https://www.linkedin.com/in/rongzhi-xu-79b050117/), Yikang Wang (https://www.linkedin.com/in/yikang-wang/). (PDF (../../static/b-15-dt.pdf))

Classification

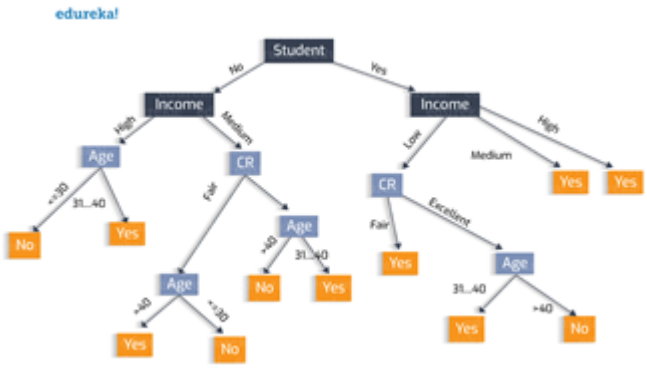
We learned about two main ways of classification: decision trees and statistical pattern recognition.

Decision Trees

Decision trees are a simple and intuitive classification model that use explicit rules to partition the feature space. In feature space, we can visually represent these partitions with vertical and horizontal splits:



For a simple decision tree example, consider a classification dataset in which we aim to classify individuals “yes” or “no” depending on whether they will purchase a computer. An example decision tree model to solve this problem is shown below:



Decision boundaries:

- Explicit: using the original features to set the boundaries
- Implicit: Partitioning via discriminant functions (using transformed features)

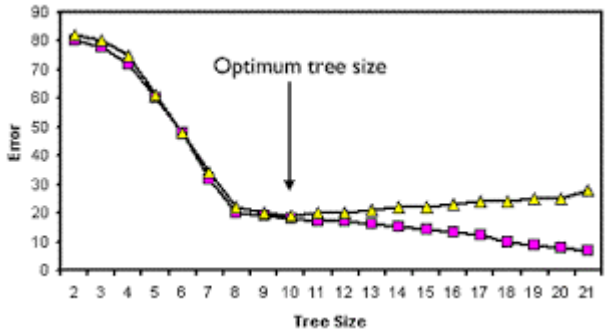
Determining which splits to make first: The solution above asserts the first split to be on whether the individual is a student. When determining which splits to make first, we would like to choose splits that are the most efficient, i.e. splits that reduce the disorder by the highest amount possible. There are two particularly popular methods to measure this disorder:

Impurity Criterion

Gini Index	Entropy
$I_G = 1 - \sum_{j=1}^{c} p_j^2$	$I_H = -\sum_{j=1}^{c} p_j log_2(p_j)$

Both are metrics of heterogeneity and aim to split the data into more homogeneous subsets. In practice, these two metrics often result in similar models and are often used interchangeably.

Determining how many splits to make: Our main goal with decision trees is to obtain a small, shallow tree with low uncertainty at the splits. While you could obtain a 100% classification rate on training data with enough splits, this will cause overfitting and lead to higher test error. Thus, it is important to validate the model to determine the number of splits that will give the highest test accuracy. This process is often called “growing and pruning.”



Tree Size: Overfitting is a significant issue for the decision tree model and it is essential to find the right size of the tree.

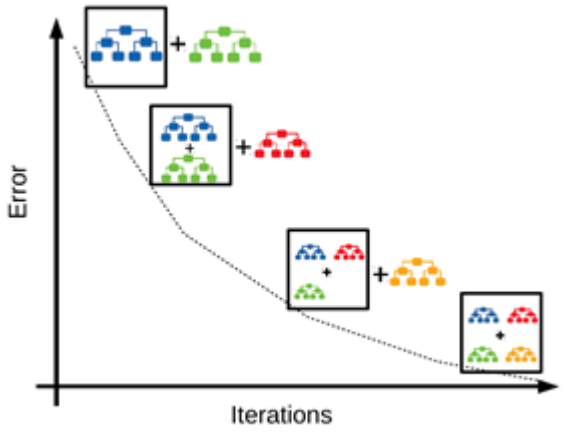
- Grow big: use a greedy, recursive forward search to build a big tree.
  - Start the tree that is a single node, then search over all possible decision rule to find the one that gives the biggest decrease in loss (increase in fit).
- Prune Back: decrease the size of decision tree.
  - Pre-Pruning: stop growing the tree earlier, before the biggest decrease in loss.
  - Post-Pruning: allows the tree to perfectly grow, and then prune back the nodes of the big tree.

Expanding Upon the Decision Tree

The decision tree model on its own is easily understood, but often provides lower classification accuracy than desired. However, there some popular expansions on the decision tree model that can lead to more robust and accurate models:

- **Random Forest:** This method essentially compounds many different decision trees in the theory that in unison, they will perform better than any single model on its own.
- **Gradient Boosting:** Another method of improving the performance of a single decision tree, gradient boosting builds on smaller trees as opposed to the fully fleshed out trees in random forests, adding one

classifier at a time to improve the model in each step forward. Essentially, gradient boosting continually adds more trees while random forest synthesizes many trees at once.



Bayes Decision Theory

Bayes' Theorem is fundamental to statistical analysis, and it's considered the ideal pattern classifier because its decision rule **automatically minimizes its loss function.**

A Review of Bayes' Theorem

Here's Bayes' Theorem:

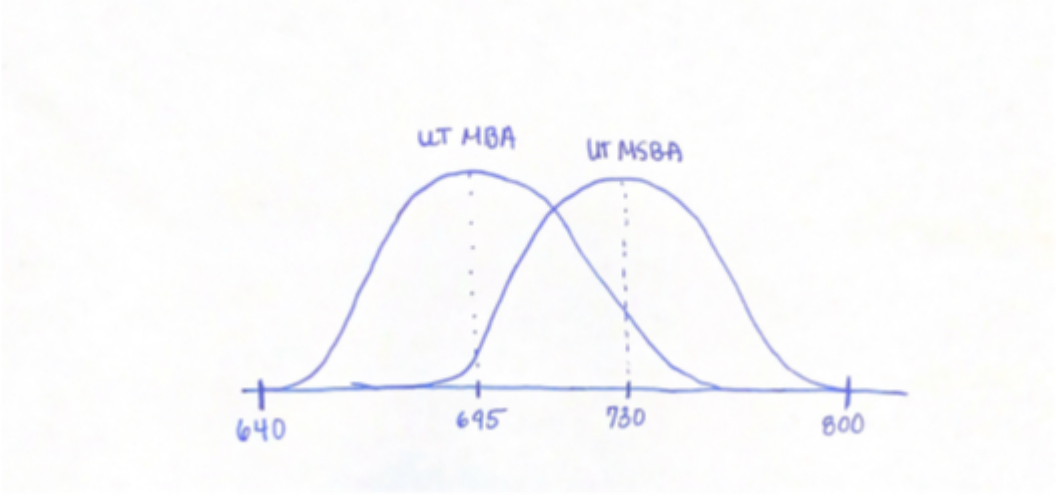
$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} = posterior = \frac{likelihood * prior}{evidence}$$

You've probably seen it many times before. But what does it really mean to us for classification problems?

Example

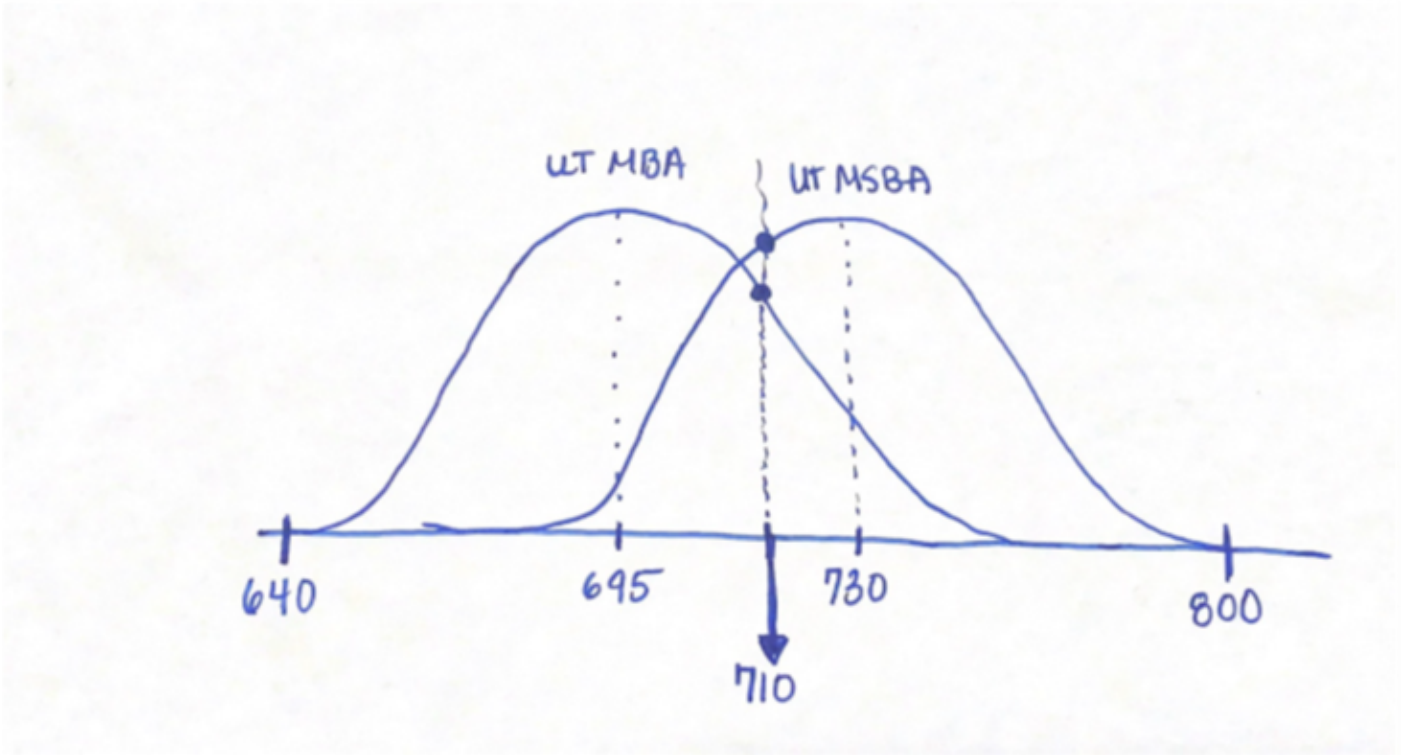
Let's put this in terms of a real-world example. Let's assume we have some data about GMAT scores and are attempting to classify students as either UT MBA students or UT MSBA students.

Here's our probability distribution for each class.



Let's assume we're given a student with a GMAT of 710, and we want to predict whether they're an MSBA or an MBA student. To do this, we need to calculate two conditional probabilities:

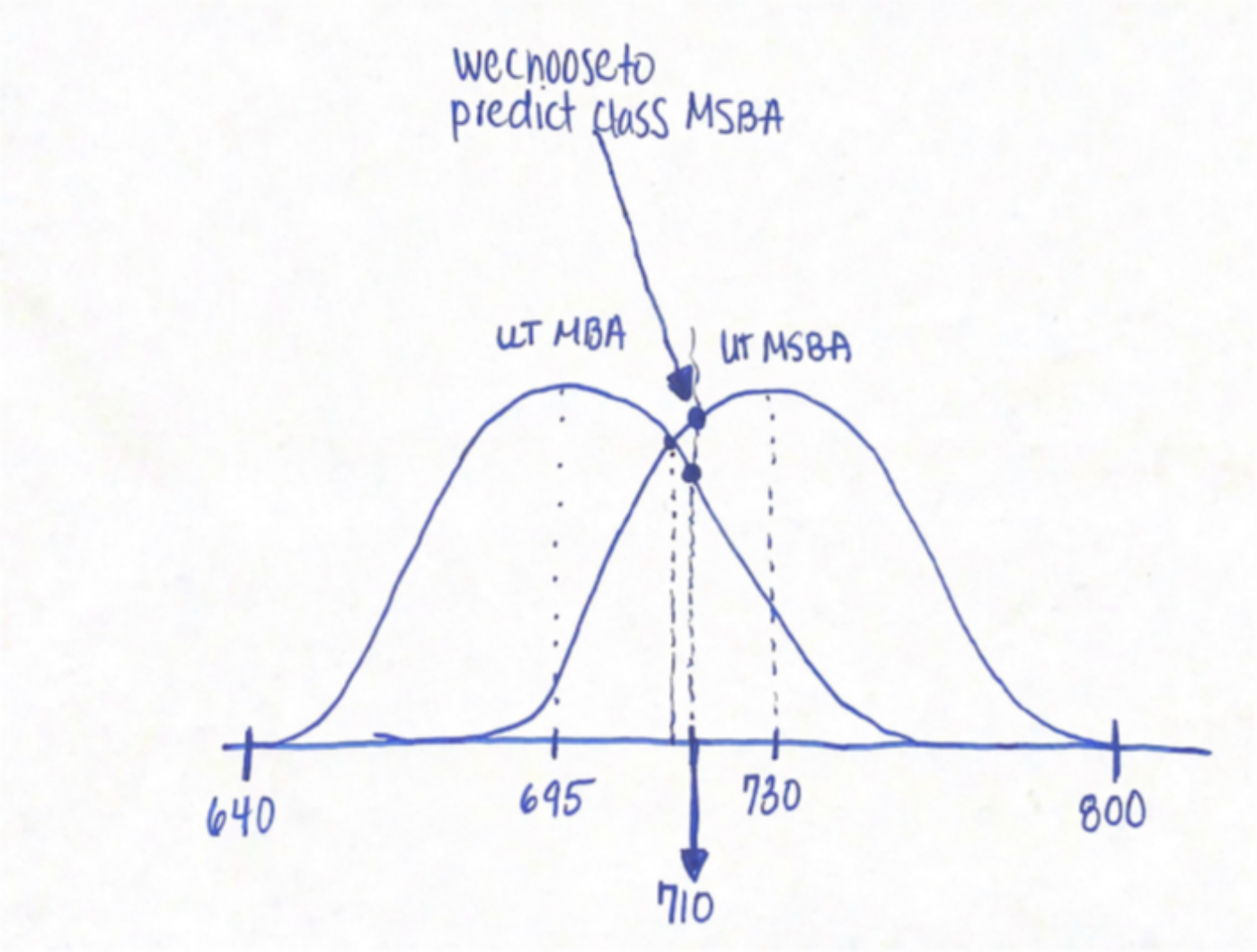
$$p(MSBA|GMAT = 710) = \frac{p(GMAT = 710|MSBA) * p(MSBA)}{p(GMAT = 710)}$$
$$p(MBA|GMAT = 710) = \frac{p(GMAT = 710|MBA) * p(MSBA)}{p(GMAT = 710)}$$



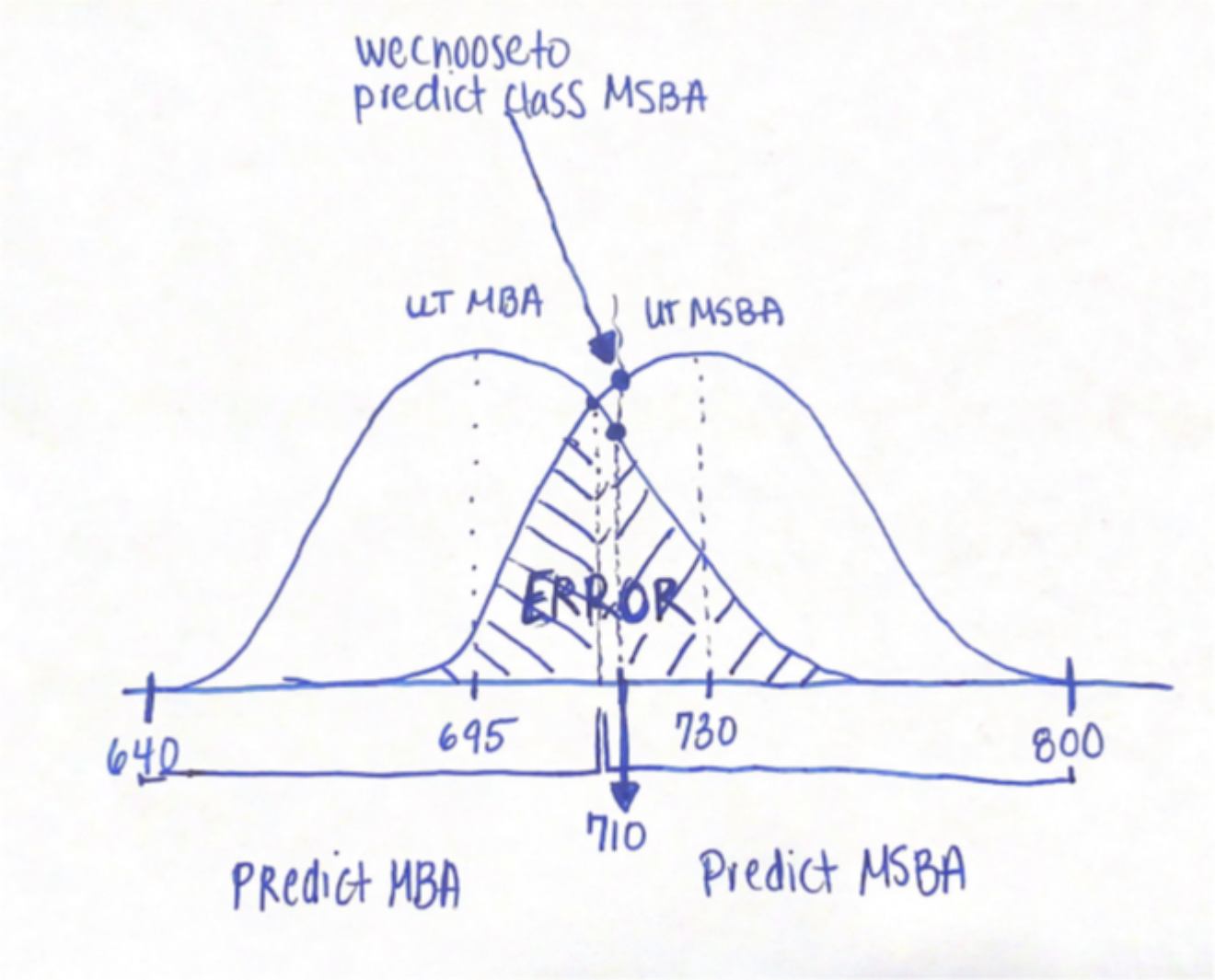
So the question is, given a GMAT of 710, which class do we predict?

The theoretical answer to this is in Bayes Decision Rule, which states that: $i = \underset{j=1...K}{argmax}\{P(C_j|x)\}$ where K is the number of classes.

Thus, in our example, we assign class K to GMAT score of 710 to whichever class K has the highest value of $p(K|GMAT = 710)$ From the image below, you see that at GMAT = 710, the MSBA class has the higher probability.



We see the irreducible error from this theory in the image below.



So **why is this optimal?** If each class is chosen to that condition risk for *each input* is minimized, then *overall risk* is minimized too.

However, it's important to remember the **assumptions** of this Bayesian model, which might not always hold:

- Decision problem must be posed in probabilitistic terms
- All relevant probability values are known

Additional Resources

- Bayesian Decision Theory - Mathematical Explanation (https://www.projectrhea.org/rhea/index.php/Bayesian_Decision_Theory)
- Bayesian Decision Theory - Graphical Explanation (https://towardsdatascience.com/bayesian-decision-theory-81103a68978e)