NON-PROBABILISTIC MODELS FOR CLASSIFICATION

LECTURES
SECTION 1
JUNE 9th

INSTITUTE FOR APPLIED
COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

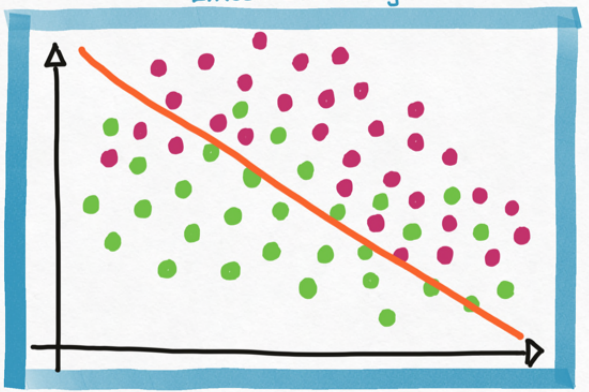
UNIVERSITY of
RWANDA

DECISION TREES

INTERPRETING CLASSIFIERS:

Logistic regression models with polynomial boundaries are flexible but harder to interpret.

Linear Boundary



model for loan approval:

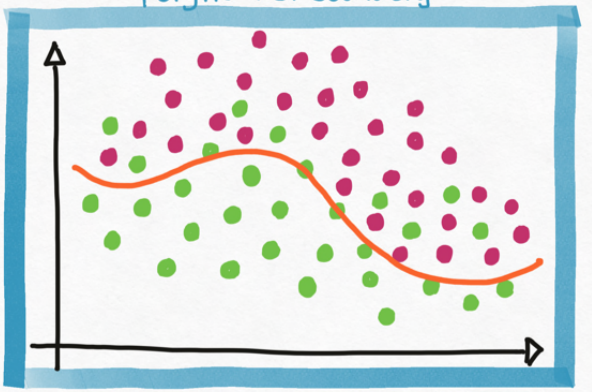
$$P(y=1|X) = \sigma(0.3X_1 + 5.1X_2 - 0.5)$$

y: approval (1), rejection (0)
X_1: income
X_2: age

In this model, age is far more important than income! The older the applicant the more likely the approval.

Should we use this model to make real life loan decisions?

Polynomial Boundary



model for loan approval:

$$P(y=1|X) = \sigma(0.3X_1^2 + 5.1X_2^2 - X_1X_2 - 0.5)$$

y: approval (1), rejection (0)
X_1: income
X_2: age

In this model, the effect of age and income are intertwined, $-X_1X_2$, and is much harder to interpret.

Should we use this model to make real life loan decisions?

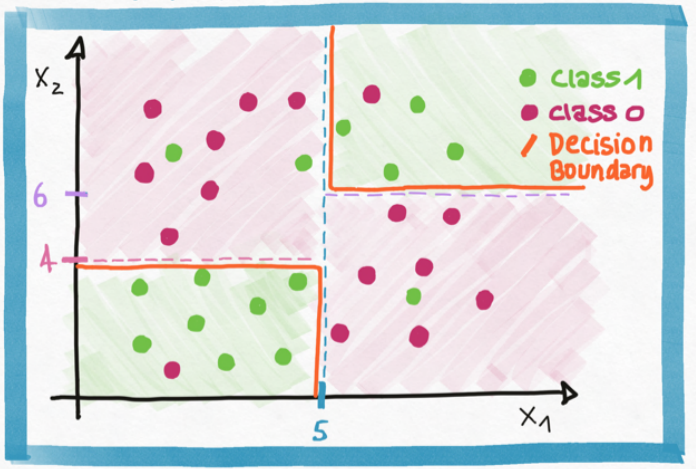
DECISION TREES:

For regression, trees (i.e. piecewise linear models) provided interpretability of linear models with the flexibility of polynomial models.

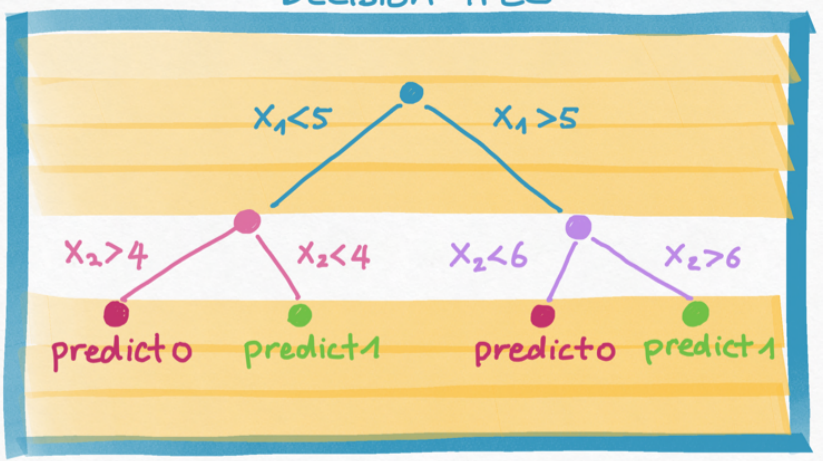
We can use a tree for classification as well: we divide up the input space into regions, we label each region based on the labels of data points in that region.

This model is called a Decision Tree.

Division of Input Space



Decision Tree

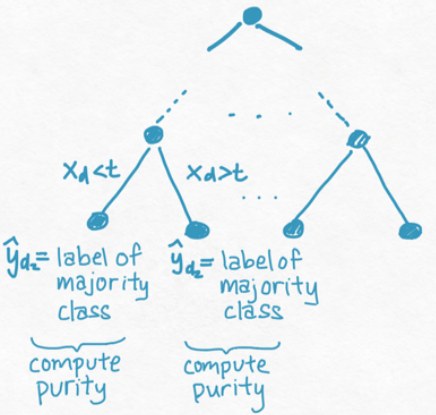


TRAINING DECISION TREES:

Given a dataset $D = \{(x^{(1)}, y^{(1)}) \dots (x^{(N)}, y^{(N)})\}$, where $x^{(n)} \in \mathbb{R}^D$ and $y^{(n)} = 0, 1$, how do we decide on the sequence of cuts to divide our domain space?

Regression tree algorithm:

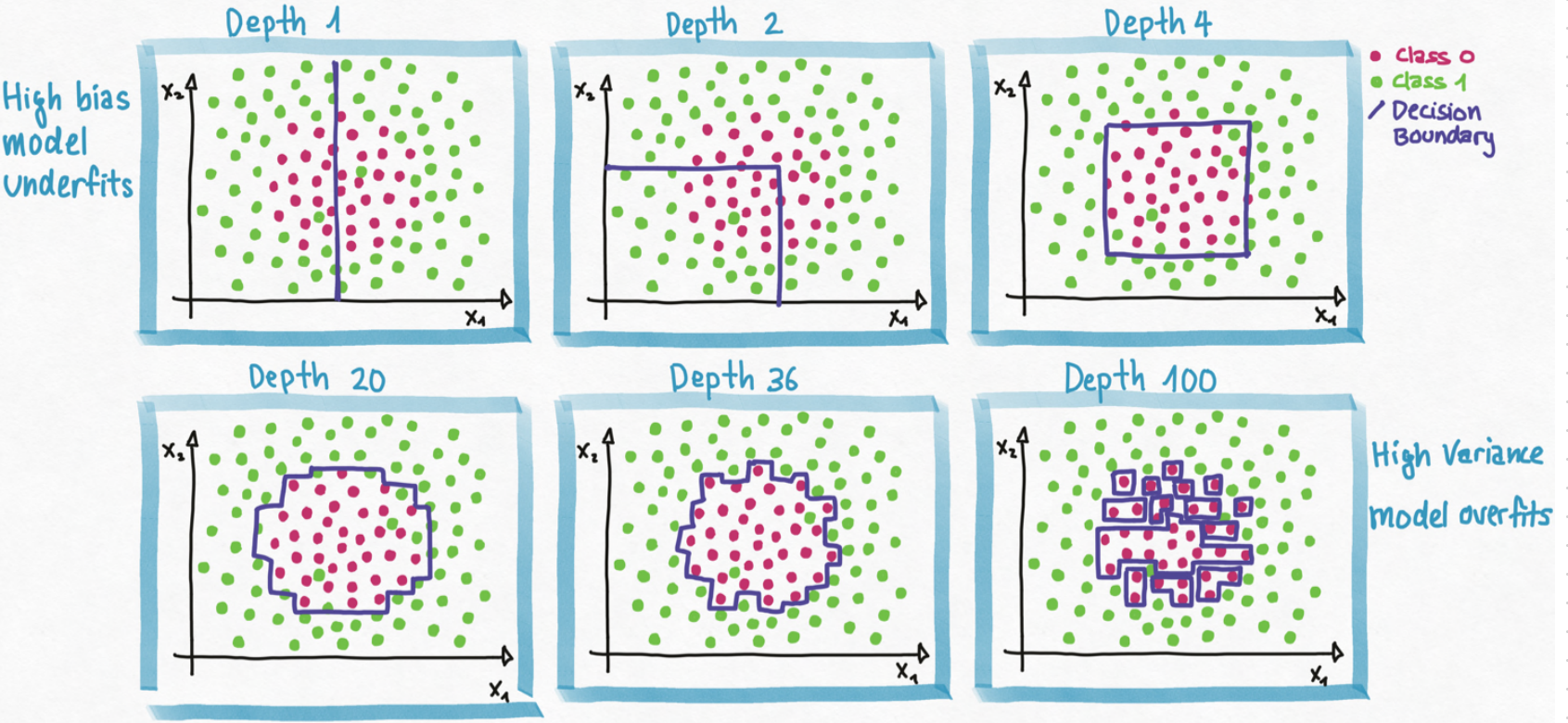
0. Start with a node containing all the data.
1. If stopping condition is not met:
   A. Choose an input dimension d and threshold t and divide the data in the node into two sets $x_d < t$ and $x_d > t$, these become new nodes.
   B. Choose d, t such that the purity of the data in each node is maximized.
2. For each new node, repeat step 1.



The stopping condition is usually a maximum depth or a minimum purity. Purity is a metric that measures the sameness of labels inside a node or region. When a region contains data almost all from one class then the purity is high.

BIAS VARIANCE TRADEOFF FOR TREES:

We've seen before that when the tree is too shallow, it cannot divide the input space into enough regions, so the model underfits. When the tree is too deep it cuts the input space into too many regions and fit to the noise of the data (overfits).

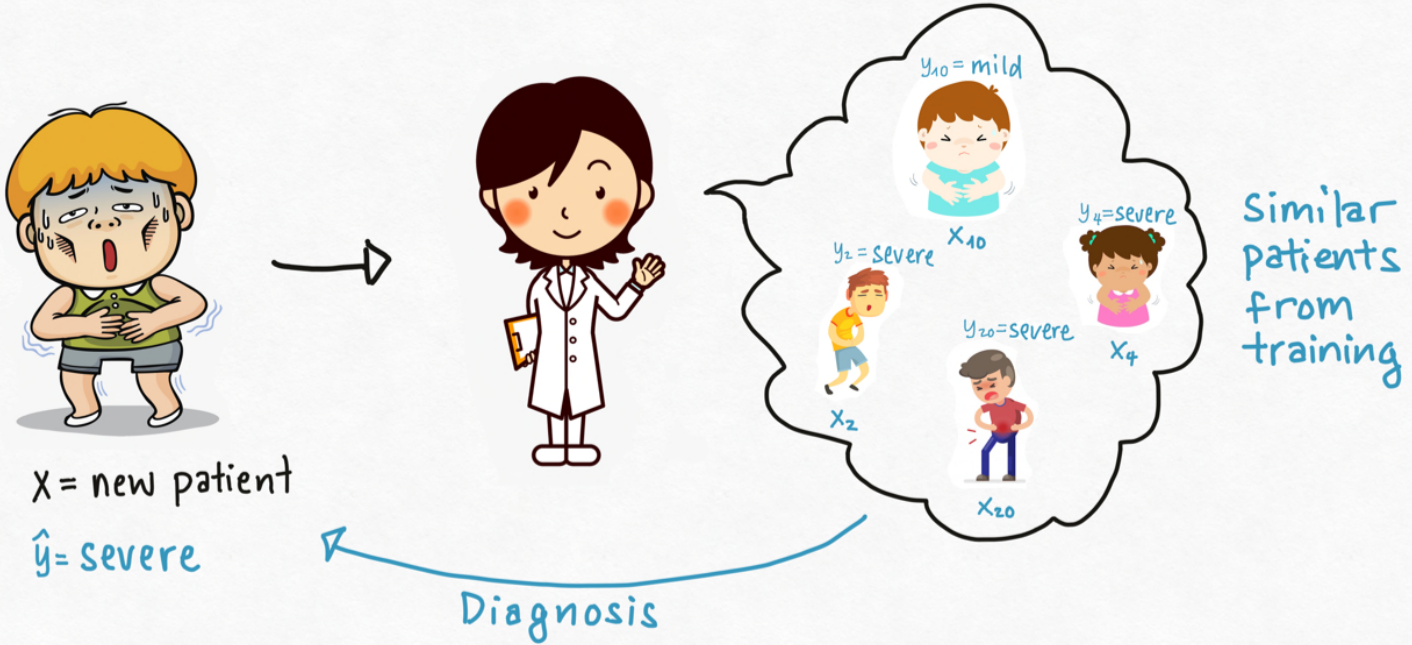


K-NEAREST NEIGHBOUR

CLASSIFIER

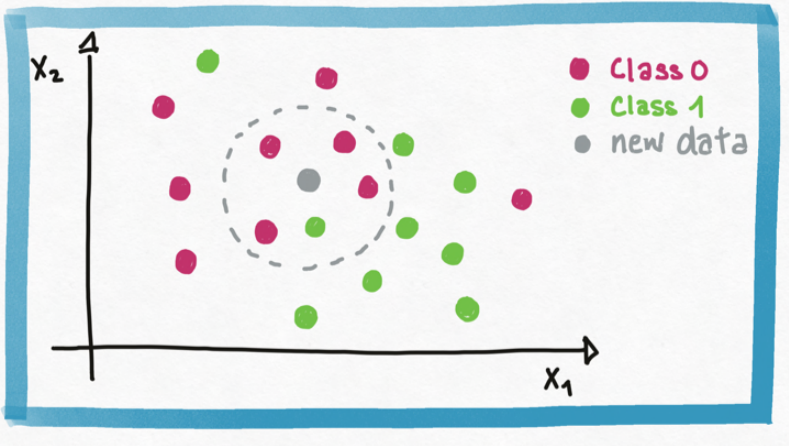
CLASSIFICATION BY EXAMPLE:

For regression, k-Nearest Neighbors is an easily interpretable model that makes predictions based on similar data points in the training data.



K-NEAREST NEIGHBOUR CLASSIFIER:

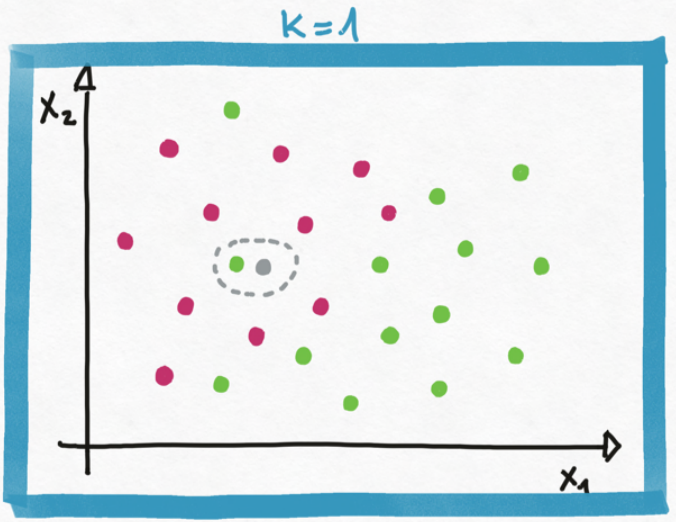
The k-nearest neighbor classifier predicts the class, $\hat{y}$, of a new data point, x, by looking at the k most similar training data points and looking for the most common class amongst these k neighbors.



Using 5 nearest neighbors for the new data point, we predict class 0 because 4 out of the 5 nearest neighbours are classified as class 0.

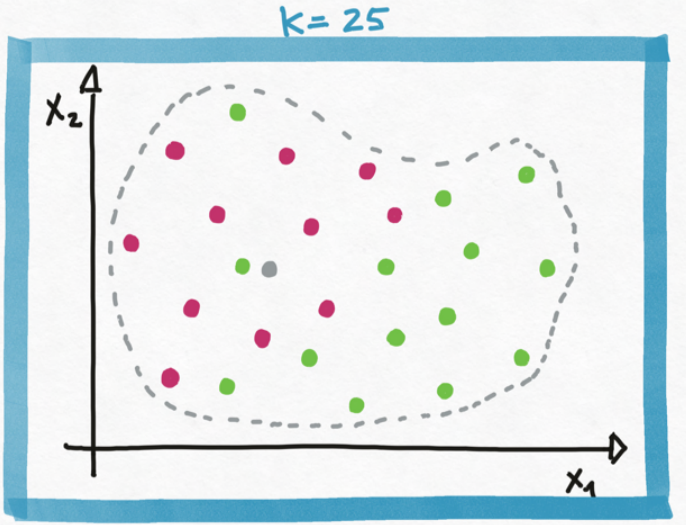
THE BIAS VARIANCE TRADE-OFF FOR KNN:

The bias and variance of KNN classifiers depend on the choice of K.



When k is small, our predictions are sensitive to the noise in the data. The model overfits.

High variance.



When k is too large, our predictions are insensitive to all local variations in the data. The model underfits.

High bias.