Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

INTRODUCTION TO BIG DATA

Juan D. Velásquez
Felipe E. Vildoso

Ingeniería Industrial

FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

CHAPTER 2

Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

LECTURE 9

Chapter 2

Ingeniería Industrial

FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

MACHINE LEARNING

MACHINE LEARNING

When you have the data in place, it's time to **extract the coveted insights**.

> *"Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed."*

Arthur Samuel, 1959

The definition leaves you with the question of how the computer learns…

MACHINE LEARNING

To achieve machine learning, experts develop **general-purpose algorithms** that can be used on large classes of learning problems.

You only need to **feed the algorithm more specific data**.

In most cases a computer will use data as its source of information and compare its output to a desired output and then correct for it.

MACHINE LEARNING

The more data or "experience" the computer gets, the better it becomes at its designated job

For example, as a user writes more text messages on a phone, the phone learns more about the messages' common vocabulary and can predict (autocomplete) their words faster and more accurately.

MACHINE LEARNING

World War II

Before

- Everything needed to be calculated by hand.
- Limited the possibilities of data analysis.

After

- Computers and scientific computing were developed.
- A single computer could do all the counting and calculations.

MACHINE LEARNING

People only need to derive the mathematical formulas, write them in an algorithm, and load their data, **BUT…**

Old algorithms didn't scale well

With the amount of data we need to analyze today, **this becomes problematic**, and **specialized frameworks** and **libraries** are required to deal with this amount of data.

MACHINE LEARNING TOOLS

Python

- Scikit-learn.

A great machine-learning toolbox and the most popular machine-learning library for Python.

- PyBrain for neural networks.

Neural networks are learning algorithms that mimic the human brain in learning mechanics and complexity. Neural networks are often regarded as advanced and black box.

MACHINE LEARNING TOOLS

- NLTK or Natural Language Toolkit

It's an extensive library that comes bundled with a number of text corpuses to help you model your own data.

- Spark

It's a new Apache-licensed machine-learning engine, specializing in real-learn-time machine learning.

Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

TYPES OF MACHINE LEARNING

SUPERVISED LEARNING

Supervised learning techniques attempt to discern results and learn by trying to find patterns in a **labeled data set**. Human interaction is required to label the data.

Ex.: Handwriting or Speech recognition.

Definition

- **Labeled Data:** Data with a category or a real-value number assigned to it that represents the outcome of previous observations.

UNSUPERVISED LEARNING

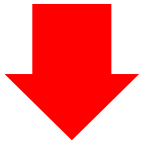
The **input data isn't labelled and the output isn't known**. The idea is to find the hidden structures within the data

Ex.: Amazon book recommendations, Google's neural network.

MACHINE LEARNING VS. FIRST PRINCIPLE

Faster implementation
Better results

Easy to generalize
Explains causality



Hard to generalize
Doesn't explain causality

Needs theory
Difficult implemetation

Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

APPLICATIONS FOR MACHINE LEARNING IN DATA SCIENCE

APPLICATIONS

- Finding oil fields, gold mines, or archeological sites based on existing sites. **(classification - regression)**
- Finding place names or persons in text. **(classification)**
- Identifying people based on pictures or voice recordings. **(classification)**
- Recognizing birds based on their whistle. **(classification)**

APPLICATIONS

- Identifying profitable customers. **(regression - classification)**
- Proactively identifying car parts that are likely to fail. **(regression)**
- Identifying tumors and diseases. **(classification)**
- Predicting the amount of money a person will spend on product X. **(regression)**

APPLICATIONS

- Predicting the number of eruptions of a volcano in a period. **(regression)**
- Predicting your company's yearly revenue. **(regression)**
- Predicting which team will win the Champions League in soccer. **(classification)**

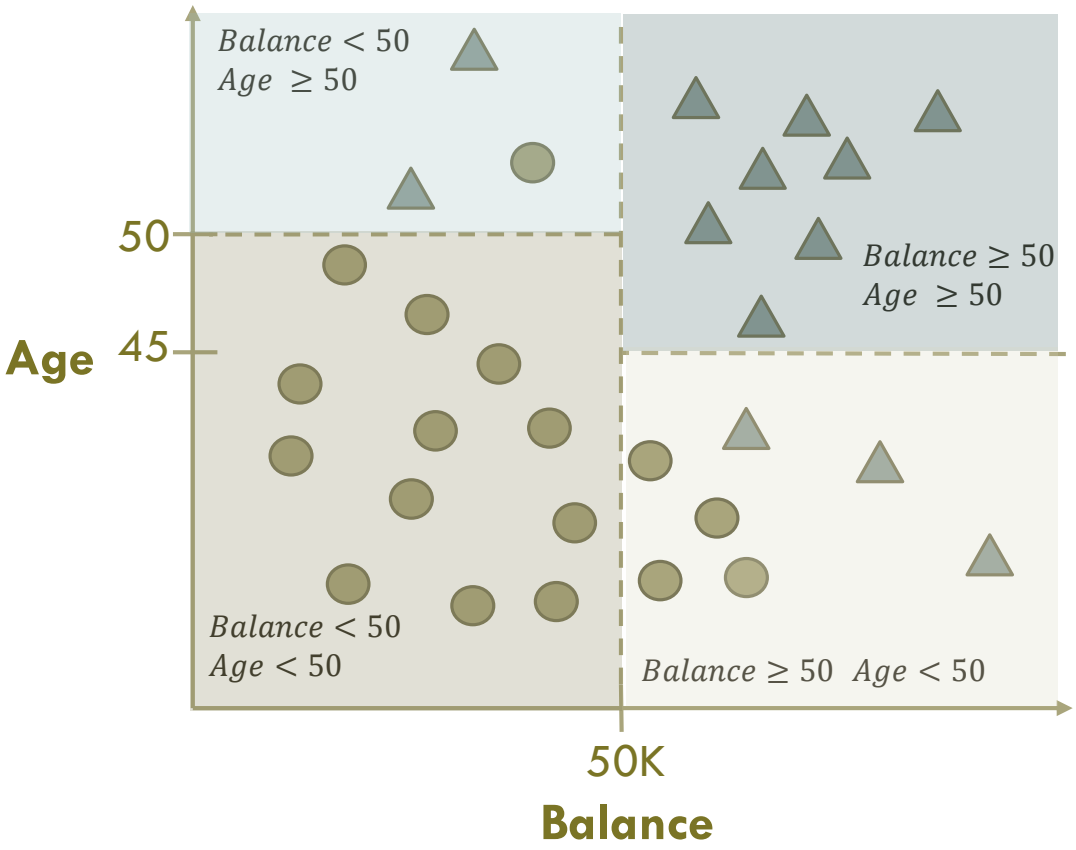
CLASSIFICATION AND REGRESSION

Regression and classification are of primary importance to a data scientist.

The **distinction** between classification and regression is whether the **value for the target variable** is **categorical** (classification) or **numeric** (regression).

Logistic regression is estimating the probability of class membership (a numeric quantity) over a categorical class. So, it's a class probability estimation model and not a regression model.

DECISION BOUNDARIES



CLASSIFICATION FUNCTION

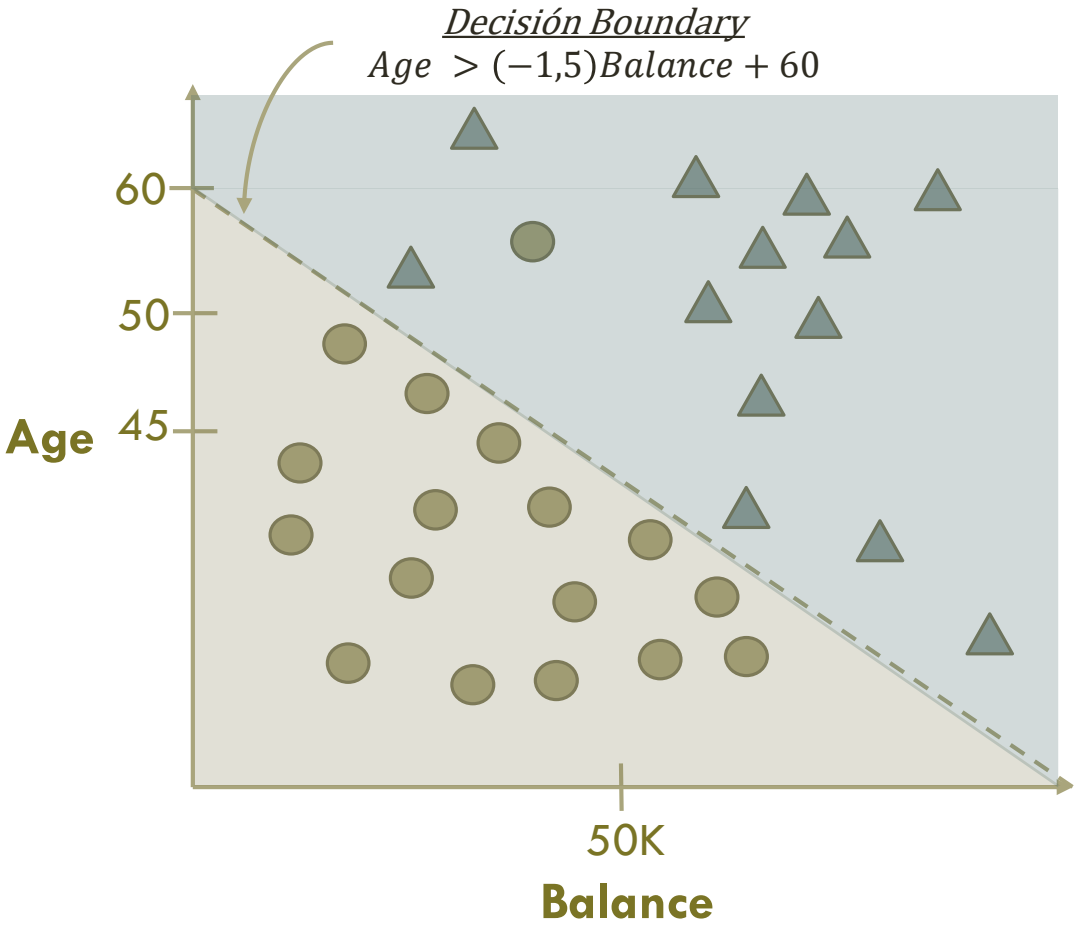
Linear discriminant:

$$class(x) = \begin{cases} + & \text{if } 1.0 \times \text{Age} - 1.5 \times \text{B}alance + 60 > 0 \\ \bullet & \text{if } 1.0 \times \text{Age} - 1.5 \times \text{B}alance + 60 \leq 0 \end{cases}$$

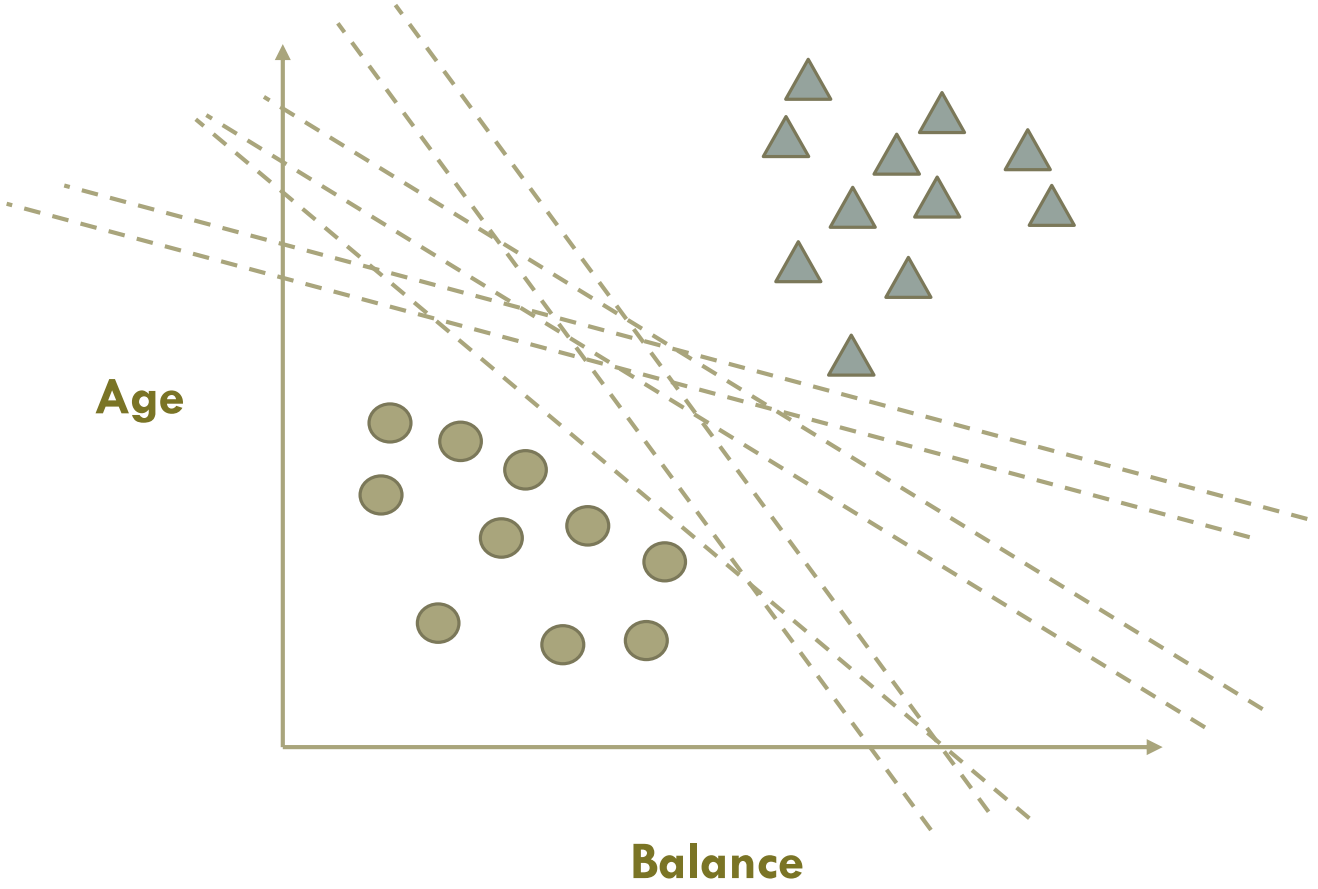
We now have a parameterized model: the weights of the linear function are the **parameters**.

The weights are often loosely interpreted as importance indicators of the features.

LINEAR CLASSIFIER



CHOOSING THE "BEST" LINE



CHOOSING THE "BEST" LINE

"Best" line depends on the **objective (loss) function**.

A loss function determines how much penalty should be assigned to an instance based on the error in the model's predicted value

OBJECTIVE FUNCTIONS

Examples of objective (or loss) functions:

- $\lambda(y;x) = |y - f(x)|$
- $\lambda(y;x) = (y - f(x))^2$ [convenient mathematically – linear regression]
- $\lambda(y;x) = I(y \neq f(x))$

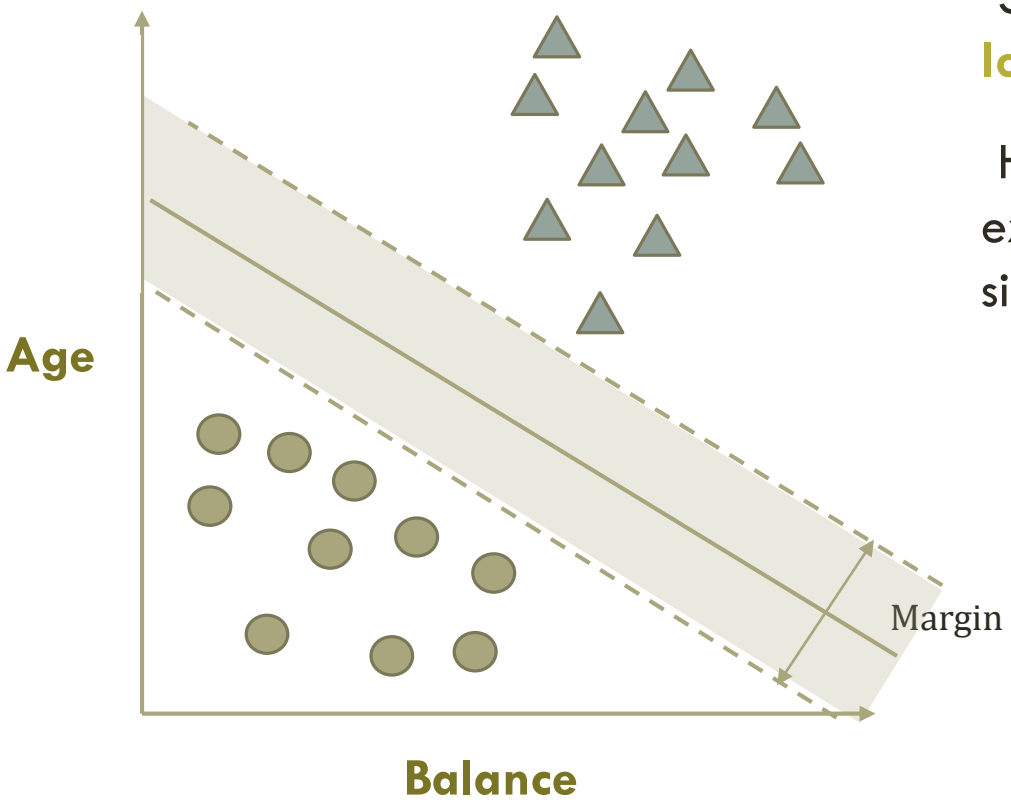
Linear regression, logistic regression, and **support vector machines** are all very similar instances of our basic fundamental technique:

> The key difference is that each uses **a different objective function**

SUPPORT VECTOR MACHINES (SVMS)

- Linear Discriminants
- Effective
- Use “hinge loss”
- Also, non-linear SVMs

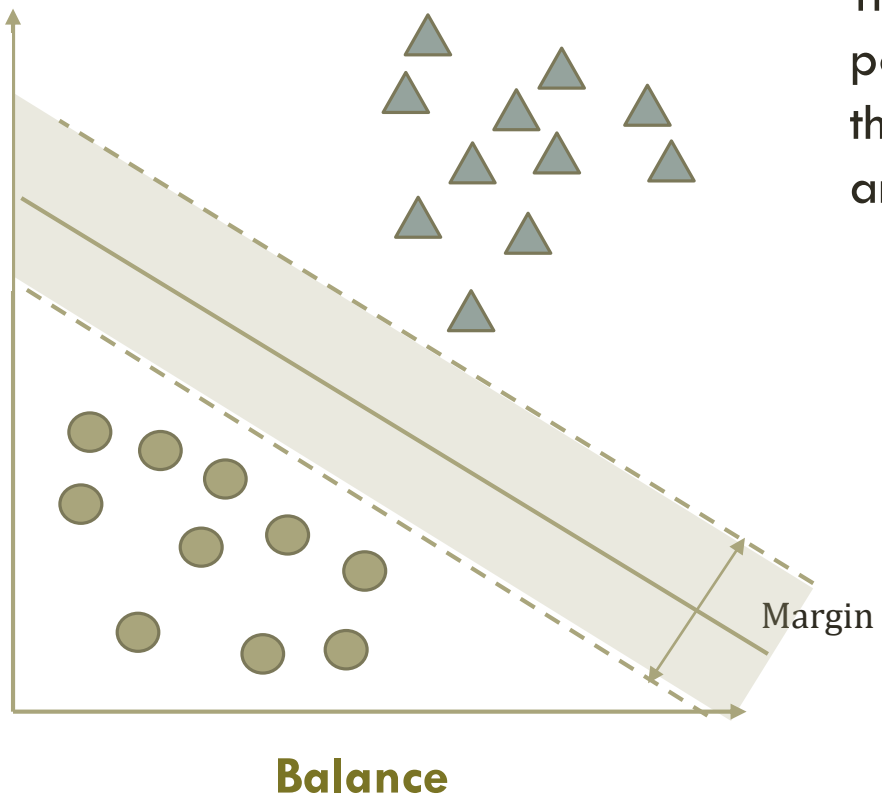
SUPPORT VECTOR MACHINES (SVMS)



Support vector machines use **hinge loss**.

Hinge loss incurs no penalty for an example that is not on the wrong side of the margin

SUPPORT VECTOR MACHINES (SVMS)



The hinge loss only becomes positive when an example is on the wrong side of the boundary and beyond the margin

- Loss then increases linearly with the example's distance from the margin
- Penalizes points more the farther they are from the separating boundary

RANKING INSTANCES AND PROBABILITY CLASS ESTIMATION

In many applications, **we don't simply want a yes or no prediction** of whether an instance belongs to the class, but we want some notion of which examples are more or less likely to belong to the class.

- Which consumers are most likely to respond to this offer?
- Which customers are most likely to leave when their contracts expire?

RANKING INSTANCES AND PROBABILITY CLASS ESTIMATION

Ranking

Business context determines the number of actions ("how far down the list")

- Tree induction
- Linear discriminant functions (e.g., linear regressions, logistic regressions, SVMs)
  - Ranking is free

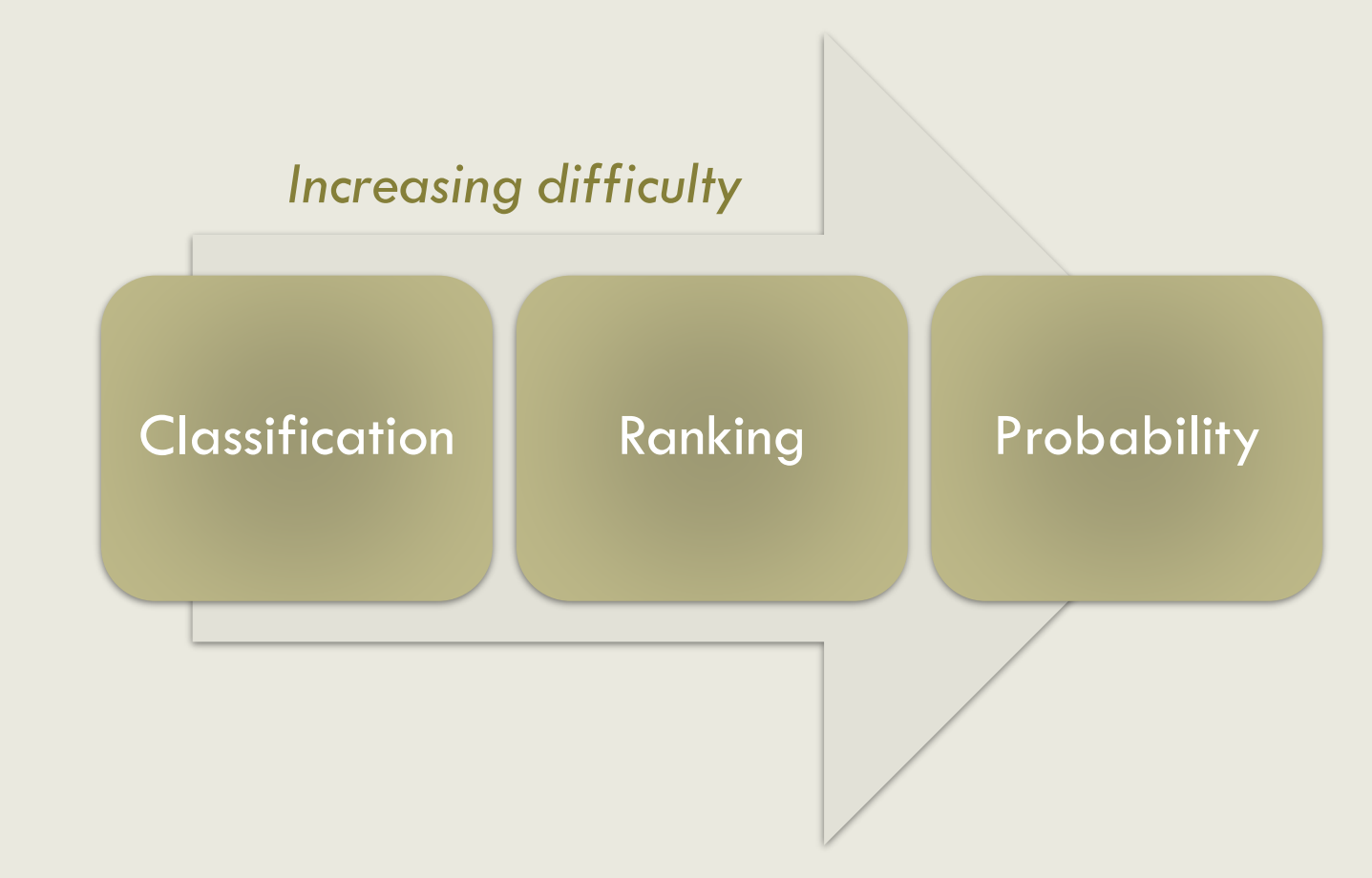
RANKING INSTANCES AND PROBABILITY CLASS ESTIMATION

Class Probability Estimation

You can always rank / classify if you have probabilities!

- Tree induction
- Logistic regression

THE MANY FACES OF CLASSIFICATION



Ingeniería Industrial
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

DECISION ANALYTIC THINKING

Evaluating Classifiers

EVALUATING CLASSIFIERS: PLAIN ACCURACY

How do we measure generalization performance?

$$Accuracy = \frac{\text{Number of correct decisions made}}{\text{Total number of decisions made}}$$

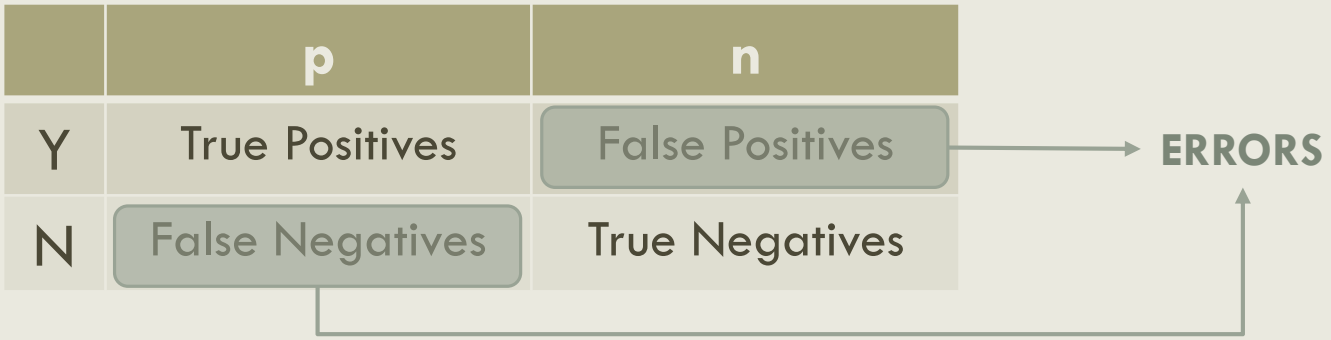
$$= 1 - \text{error rate}$$

Too simplistic...

EVALUATING CLASSIFIERS: THE CONFUSION MATRIX

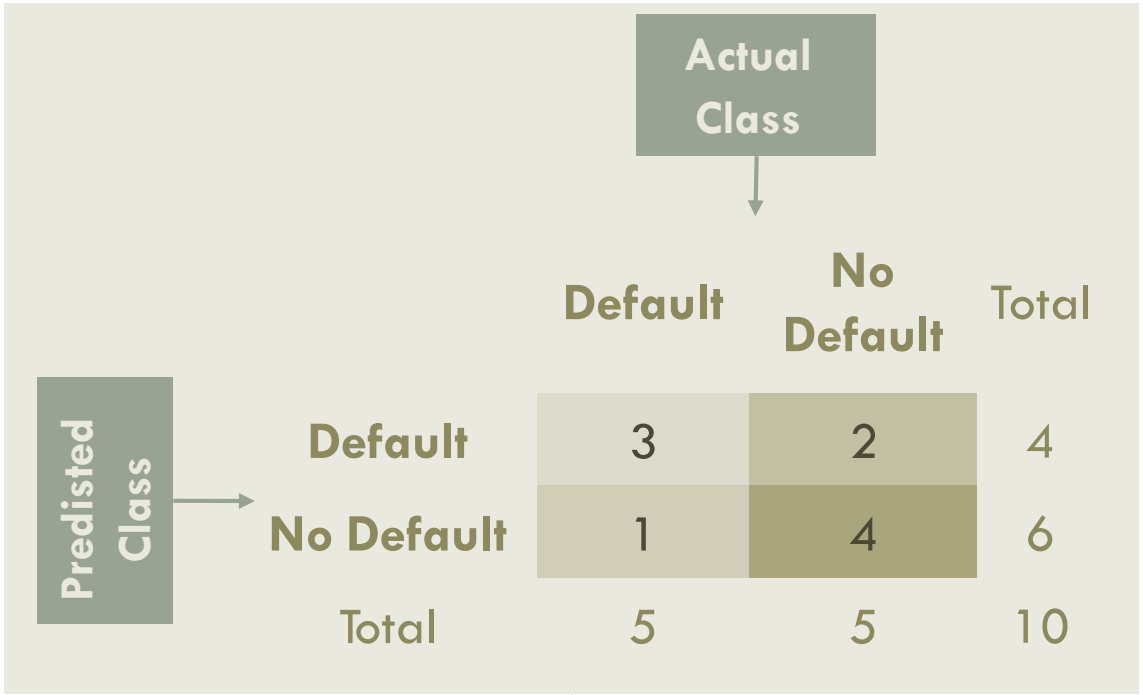
A confusion matrix for a problem involving n classes is an n × n matrix that separates out the decisions made by the classifier.

- Columns: actual classes.
- Rows: predicted classes.



BUILDING A CONFUSION MATRIX

Default Truth	Model Prediction
0	0
1	1
0	1
0	1
0	0
1	1
0	0
0	0
1	1
1	0



OTHER EVALUATION METRICS

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F - measure = 2\left(\frac{Precision\ x\ Recall}{Precision + Recall}\right)$$

Ingeniería Industrial

FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

QUESTIONS?

SEE YOU ON THE NEXT CLASS!