Data Science for Business
Lecture #5
Predictive Models

Prof. Alan L. Montgomery

The University of Hong Kong & Carnegie Mellon University, Tepper School of Business

email: alanmontgomery@cmu.edu

What is a predictive model?

What is a model?

A *simplified* representation of reality created for a specific purpose.

- Simplification is based upon some assumptions, which may not be entirely accurate. Inherent tradeoff between complexity and accuracy.

Examples:

- Map, Prototype, Black-Scholes model

Data Mining Examples:

- "formula" for predicting probability of customer attrition at contract expiration
- "classification model" or "class-probability estimation model"

Which model is best?

What is a predictive model?

Use a set of inputs (or features|covariates|attributes|independent variables) to forecast an output (or target|dependent)

- It is an *abstraction*. Abstractions are meant to capture the most important elements – they are not perfect.
- They are *mathematical* models. We quantify the relationships between the inputs and outputs
- They are not perfect, so usually we think in *probabilistic* terms and use *statistical* techniques

Optimal Pricing

- *Inputs*: Vector of all prices and past purchases
- *Output*: Category profit
- *Model*: Log-linear regression

Customer Churn

- *Inputs*: Vector of customer characteristics (tenure, usage, …)
- *Output*: Probability of churn next month
- *Model*: Logistic regression

Product Recommendation

- *Inputs*: Vector of all past purchases for a customer
- *Output*: Vector of purchase probabilities for all new products
- *Model*: k-nearest neighbor

Pattern/Model?

NAME	Balance	Age	Default
Mike	123,000	30	yes
Mary	51,100	40	yes
Bill	68,000	55	no
Jim	74,000	46	no
Mark	23,000	47	yes
Anne	100,000	49	no

Pattern 4:

If **Balance** >= 50K and **Age** > 45
Then **Default** = 'no'
Else **Default** = 'yes'

Good vs bad patterns?

Pattern 1:

If **Names starts with M**
Then **Default** = 'yes'
Else **Default** = 'no'

Pattern 2:

Age is inversely proportional to alphabetical order

Pattern 3:

Young people are more likely to default

Pattern 5:

If **Names ends with 'e'**
Then **Balance > 100000**
Else **Balance** <100000

Discussion Exercise

A large bank wishes to target its existing customers to take out a new credit card using its various advertising channels: emails, web messages on its website, ATM splash screens, and direct mail. The bank wants to design an optimal advertising campaign to profitably recruit new credit card customers. The bank is able to use its internal customer mart about existing customers.

Questions:

- *Define the inputs and outputs for the model.* Consider the internal information collected from past customer transactions, as well as external information from credit bureaus like Equifax and marketing research firms like Nielsen.
- *What level of abstraction is best?* Should we model everyone customer or groups? Should we model daily or monthly response?
- *What relationships do you expect to exist between each input and output?* Linear or Non-linear?
- *Formulate a mathematical model.* What is the association between the inputs with the outputs.
- *How do we connect the model with the decision?* How would you use your model to decide which messages to give a customer?

Modeling Terminology

Supervised Data Mining: Terminology

Example, Instance

A fact; a data point

Attributes/Features

One example

Name	Balance	Age	Default
Mike	123,000	30	yes
Mary	51,100	40	yes
Bill	68,000	55	no
Jim	74,000	46	no
Mark	23,000	47	yes
Anne	100,000	49	no

A data set/ sample *(as noun)*

A set of examples

"To sample": to choose certain examples

an example of this form sometimes is called a "**feature vector**"

Feature Types

Numeric: anything that has some order

- Numbers (that mean numbers)
- Dates (that look like numbers …)
- **Dimension** of 1

Categorical: stuff that does not have an order

- Binary
- Text
- **Dimension** = number of possible values (minus 1)

What type of data? Moody's Ratings, Industry codes

Dimensionality of the data?

Attributes/Features

Name	Balance	Age	Default
Mike	123,000	30	yes
Mary	51,100	40	yes
Bill	68,000	55	no
Jim	74,000	46	no
Mark	23,000	47	yes
Anne	100,000	49	no

"Dimensionality" of a dataset is the sum of the number of numeric features and the number of values of categorical features

Supervised Data Mining: Terminology

Example, Instance

A fact; a data point

typically described by a set of attributes (fields, variables, features) and a target variable (label).

Attributes → Target

One example →

Name	Balance	Age	Default
Mike	123,000	30	yes
Mary	51,100	40	yes
Bill	68,000	55	no
Jim	74,000	46	no
Mark	23,000	47	yes
Anne	100,000	49	no

Equivalent statistical terminology :

Attributes - independent variables

Target - dependent variable

Dimensionality: sum of dimensionality of the attributes excluding target

Data Mining : Basic Terminology

Training (a.k.a. ***learning***, ***induction, inductive learning, model induction, estimation***)

A process by which a pattern/model is generalized from factual data

NAME	Balance	Age	Default
Mike	123,000	30	yes
Mary	51,100	40	yes
Bill	68,000	55	no
Jim	74,000	46	no
Mark	23,000	47	yes
Anne	100,000	49	no

Data Mining: Terminology

A learner (inducer, induction algorithm, estimator)

A method or algorithm used to generalize a model or pattern from a set of examples

Name	Balance	Age	Default
Mike	123,000	30	yes
Mary	51,100	40	yes
Bill	68,000	55	no
Jim	74,000	46	no
Mark	23,000	47	yes
Anne	100,000	49	no

Learner:
Induces a model from examples

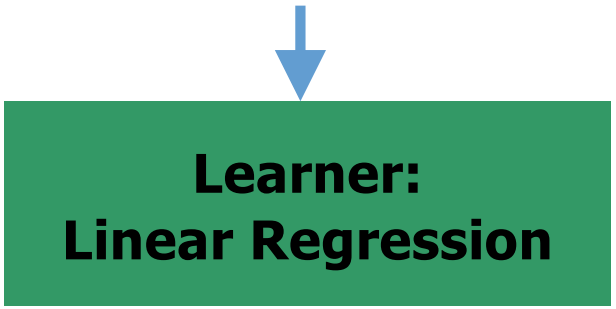
Classification Model:

```
If Balance >= 50K and Age > 45
   Then Default = 'no'
   Else Default = 'yes'
```

Data Mining: Terminology
Regression modeling (rather than classification modeling)

Target Variable

Name	Income	Age	Order $ Amount
Mike	123,000	30	183
Mary	51,100	40	131
Bill	68,000	55	178
Jim	74,000	46	166
Mark	23,000	47	117
Anne	100,000	49	198



Model:

Amount= 0.001*__Income__+2*__Age__

Conclusions

Predictive Modeling

Predictive modeling introduces the notion of a mathematical model that relates inputs to outputs

The steps in the modeling process

- Selecting the model
- Training/Estimating the parameters
- Validating the model
- Usage of the model