Using Analytic Solver Platform
What We’ll Cover Today

• Introduction
  • Session III beta training program goals
  • Brief overview of XLMiner and what we have learned
• Supervised learning – prediction
• Unsupervised learning – association rules
• Time series forecasting – smoothing
Session III Online Beta Training Goals

• To empower you to achieve success
  • State of the art tools
  • Online educational training
  • Training documents and demos

• To familiarize you with the following concepts:
  • Understanding the ideas behind the prediction techniques
  • Fitting prediction models to data
  • Assessing the performance of methods
  • Applying the models to predict unseen test cases
  • Using affinity analysis
  • Forecasting time series using smoothing techniques
Brief Overview of XLMiner

Data Analysis

• Draw a sample of data from a spreadsheet, or from external database (MS-Access, SQL Server, Oracle, PowerPivot)
• Explore your data, identify outliers, verify the accuracy, and completeness of the data
• Transform your data, define appropriate way to represent variables, find the simplest way to convey maximum useful information
• Identify relationships between observations, segment observations
Brief Overview of XLMiner

- Analytic Solver Platform's XLMiner component offers over 30 different methods for analyzing a dataset to gain new insights.

**Time Series**
- Forecast the future values of a time series from current and past values
- Smooth out the variations to reveal underlying trends in data
  - Economic and business planning
  - Sales forecasting
  - Inventory and production planning
Brief Overview of XLMiner

**Data Mining**
- Partition the data so a model can be fitted and then evaluated
- Classify a categorical outcome – good/bad credit risk
- Predict a value for a continuous outcome – house prices
- Find groups of similar observations – market basket analysis
Supervised Learning Algorithms

• For each record:
  • Outcome measurement $y$ (dependent variable, response, target).
  • Vector of predictor measurements $x$ (feature vector consisting of independent variables).

• Classification:
  • Bank Customer: Loan (Yes / No)?

• Prediction:
  • Housing market: Price.
  • Product: Demand.
Unsupervised Learning Algorithms

• No outcome variable in the dataset, just a set of variables (features) measured on a set of samples.
  • Market basket analysis.
Chapter 6 – Part II
Prediction Methods

Using XLMiner
Prediction Using XLMiner

- Multiple Linear Regression
- k-Nearest Neighbors
- Regression Tree
- Neural Networks
Multiple Linear Regression (MLR)

- Fundamental and most-widely used technique for supervised learning.
- Main assumption: linear dependence of response variable on predictors.
- Despite linearity assumption, Linear regression is useful conceptually and practically.
- MLR assumes that the residuals (error terms) are normally distributed.
- Models are fitted and parameters are estimated using Least Squares approach.
- XLMiner: comprehensive toolkit for Regression Models with advanced statistics and diagnostics reports.
- XLMiner MLR: 5 embedded feature selection techniques including 4 heuristic and 1 exact algorithms.
Strengths and Weaknesses of Multiple Linear Regression

**Strengths:**

- Very often linear relationship can serve as a good approximator of real dependency.
- Has a closed-form solution.
- Least squares procedure yields optimal estimates of parameters.
- Data is used “efficiently” – MLR is able to learn from small data.
- Applicable to Big Data.
- Theory is well-developed – one can access comprehensive information to support the model.
Strengths and Weaknesses of Multiple Linear Regression

**Weaknesses:**

- Real relationship is rarely linear.
- Ordinary MLR doesn’t account for dependence between predictors.
- Results of Linear Regression analysis do not show causality.
- Sensitive to outliers.
Summary – Score Test Data

• Click **Score** on the XLMiner ribbon.

• Select the new data and the Stored Model worksheets.

• Click **Next**. XLMiner will open the *Match variables – Step 2* dialog.

• Match the Input variables to the New Data variables using **Match variable(s) with same names(s)** or **Match variables in stored model in same sequence**.

• Then click **OK**.
Summary – Multiple Linear Regression

- Select a cell on the Data_Partition1 output worksheet, then click Predict – Multiple Linear Regression.
- Choose input and output variables.
- Choose desired options and click Finish.
Prediction Using XLMiner

- Multiple Linear Regression
- k-Nearest Neighbor
- Regression Tree
- Neural Networks
k-Nearest Neighbors

- Powerful algorithm which makes prediction decisions based on information from neighboring records:
  - Identifies the k observations in the training data that are most similar to a given observation.
  - Response is predicted based on average of neighbors’ responses, weighted according to similarity.
- No fitted model parameters – training data is our model.
- Similarity measure is Euclidean Distance.
- Requires independent variables to be scaled appropriately.
- Best model can be chosen by assessing the prediction error for various values of k.
- Model should be tested on validation data to decrease chance of overfitting.
Strengths and Weaknesses of the $k$-Nearest Neighbor Algorithm

Strengths:

• Performs well in practice.
• Produces stable and easily interpretable results.

Weaknesses:

• Computationally and memory-wise expensive.
• Focuses on local structure of data, fails to capture global picture.
• “Curse of dimensionality.” In high dimensions, the concept of “nearest neighbors” becomes more and more blurry.
• Extremely sensitive to outliers and noise.
• May demonstrate poor performance on data with undersampled/oversampled groups.
Summary – *k*-Nearest Neighbor

- Select a cell on the `Data_Partition1` worksheet, then click `Predict – k-Nearest Neighbors`.
- Select desired variables under `Variables in input data` then click > to select as input variables. Select the output variable or the variable to be classified.
- Specify “Success” class and the initial cutoff value, and click `Next`.
- Select `Normalize input data` and the reports and input Number of nearest neighbors. Click `Finish`.

[Image of software interface for k-Nearest Neighbor model]

4/2/2014
Prediction Using XLMiner

- Multiple Linear Regression
- $k$-Nearest Neighbor
- Regression Tree
- Neural Networks
Regression Tree

- Partitions the space of independent variables using set of splitting rules. This process is summarized and visualized by a tree.
- Works from the root node to leaves, identifying the “best” splits according to a purity measure of observations in the child nodes.
- Each internal node corresponds to the feature used for splitting.
- Each branch leads to node’s children – defines two subsets of possible values of parent node.
- Leaf (terminal) nodes represent the value of response – given the path from the root to the terminal node.
Regression Tree

• A fully grown regression tree is very likely to overfit training data:
  • Solution: pruning – reduces the tree size by removing subtrees that provide little contribution to predictive power of the model.

• Pruning is extremely useful, as a technique to reduce overfitting, and as a method of creating simpler, more interpretable, robust models.

• However, “over-pruned” trees may lose their ability to capture structural information. What is the optimal size of a decision tree?

• There are various techniques for “optimal” pruning.
  • Main idea: reduce the size of the tree without sacrificing the predictive accuracy.
  • XLMiner: cross-validation pruning. Uses validation partition to assess the predictive error of the model.
Strengths and Weaknesses of Regression Trees

**Strengths:**

- Produces easily interpreted model.
  - Transparent results, can be interpreted as explicit if-then rules by non-expert users.
- Works well with raw data that has not been preprocessed, potentially having different scales, missing values and outliers.
- Computationally efficient for moderate size datasets.
- Implicit *feature selection*: top nodes correspond to most informative, important features according to classification tree model.
- Does not impose explicit assumptions about underlying relationships in data.
Strengths and Weaknesses of Regression Trees

Weaknesses:

• Provides only greedy heuristic approach for generally NP-Hard problems. Solution corresponds to local optimum.
• Often predictive accuracy of regression trees is weaker than other prediction techniques.
Summary – Regression Tree

- Select a cell on the Data_Partition1 worksheet, then click Predict – Regression Tree on the XLMiner ribbon.

- Select the desired options in step 2 of 3 dialog box.

- Set Maximum # levels to be displayed, select Full tree, Best pruned tree, Minimum error tree, and reports, then click Finish.

- Select Output and Input variables.
Prediction Using XLMiner

• Multiple Linear Regression
• k-Nearest Neighbor
• Regression Tree
• Neural Networks
Neural Networks

• Artificial Neural Network (ANN) is a complex learning system inspired by the structure of the human brain.

• ANN is an umbrella term for many powerful machine learning techniques.

• XLMiner - comprehensive tool for feed-forward back-propagation Neural Networks.

• ANN is a system of interconnected neurons, which are organized in layers.

• Neurons represent computational units that perform weighted averaging and “activation” of information circulating through the network.

• ANN is adaptive technique that is able to internally perform feature extraction, capturing complicated nonlinear relationships.

• Highly dependent on initial settings, architecture.
Neural Networks Key Components

- Input neurons – features.

- Next, information is forwarded deeper into the network, resulting in prediction on the output layer.

- Error is measured (training, cross-validation) and back-propagated to the network to adjust the weights – network has just learned something from training data.

- The process is repeated for each training record. Processing of all training records is one iteration or epoch.

- Perform as many learning epochs as necessary to achieve desired predictive accuracy (measure training, cross-validation errors).
Strengths and Weaknesses of Neural Networks

**Strengths:**

- “Universal Approximators”. Come to play when the nature of data is barely interpretable.
- Able to detect highly nonlinear relationships between independent and dependent variables.
- Able to detect and take into account relationships between predictors.
- Learning is automated to some extent – less formal modeling.
- Provide robust models for large high-dimensional datasets, overcoming many problems of conventional learning techniques.
- No strong explicit assumptions involved.
Strengths and Weaknesses of Neural Networks

Weaknesses:

• “Black-box” learning: models are almost not interpretable. Doesn't provide insight into the structure of the relationships.
• Computationally expensive.
• Prone to overfitting, unless necessary steps are taken to prevent it.
• Greatly depends on chosen architecture, optimization parameters, choice of activation and error functions. However:
  • General rules exist to simplify above choices.
  • XLMiner – Automatic Network Architecture option.
Summary – Neural Networks

• Select a cell on the *Data_Partition1* worksheet, then click **Predict – Neural Network**.

• Select Input and Output variables.

• Select **Normalize input data**. Manfully adjust the **Network Architecture** and **Training options**.

• Select the **Reports** and click **Finish**.
Comments on Prediction

- In the real world it is impossible to find a perfect model. Each of them may produce specific set of prediction rules, leading to different results and different predictive power and accuracy.

- Data analysts typically build several models (e.g., Multiple Linear Regression, $k$-Nearest Neighbor, Regression Trees and Neural Networks) and choose one that achieves best overall performance depending on application needs.
Comments on Prediction

• Two fundamental problems exist and should be taken care of:
  • Overfitting – models try hard to explain training data, yet fail to generalize on new incoming patterns:
    • Simple VS. Complex model – choose simple when possible.
    • Use cross-validation to test your model against unseen samples.
  • Curse of dimensionality – volume grows exponentially with number of dimensions:
    • Choose algorithm accordingly with number of dimensions.
    • Try to reduce dimension of your data (explicitly or using XLMiner’s techniques for feature selection and extraction).
  • Use test samples to provide final independent test on the model predictive power.
Affinity analysis

Using XLMiner
Association Rules

• Delivers “what goes with what” by examining if-then rules and selecting those that are most likely indicators of true dependence.

• If A then B: “if” and “then” parts are called antecedent and consequent respectively.

• Support of a rule is percentage of total number of records that include both antecedent and consequent.

• Confidence of a rule – $P\{\text{antecedent}|\text{consequent}\}$.

• Lift Ratio of a rule is a measure of usefulness – Confidence /$P\{\text{consequent}\}$.
  • Lift Ratio greater than 1 suggests usefulness.
Strengths and Weaknesses of Association Rules

Strengths:
• Generates clear, simple rules. Transparent and easy to understand.

Weaknesses:
• Rare combinations tend to be ignored since they do not meet the minimum support requirement.
  • Use higher level hierarchies as the items.
Summary – Association Rules

- Select a cell in the dataset, then click Associate – Association Rules.
- Select the *Input data format*.
- Enter desired value for the *Minimum Support (# transactions)*.
- Enter desired value for *Minimum confidence*.
- Click OK.
Time Series – Smoothing

Using XLMiner
Time Series Forecasting

- Time series is a set of observations on a quantitative variable collected at equal time intervals.
- Extrapolation models analyze the past behavior of a time series variable to forecast future.
  \[ \hat{Y}_{t+1} = f(Y_t, Y_{t-1}, Y_{t-2}, ...) \]
- XLMiner includes ARIMA and smoothing methods.
  - See recorded video of ARIMA methods.
  - This session covers exponential smoothing methods.
Time Series – Smoothing

- Smoothing techniques smooth out random variations in time series data and reveal underlying trends and patterns.
- In stationary time series, statistical properties do not change over time.
  - There is no significant upward or downward trend in data.
- Stationary – Exponential and Moving Average.
  - One-step ahead forecast is the smoothed value of the last observation.
- Trend – Double Exponential.
  - Trend is the long-term sweep or general direction of movement in a time series.
  - XLMiner includes feature optimization for the parameters.
- Trend and seasonality – Holt-Winters.
Strengths and Weaknesses of Smoothing

**Strengths:**
- It is widely used in time series data with trend and performs well.
- Easy to use.
- Applicable for short-run forecasting.

**Weaknesses:**
- Not very accurate when a longer forecasting horizon is necessary.

Note: User needs to understand the data in order to choose the right model and parameter.
Summary – Holt-Winters’ Smoothing

• Click a cell in the dataset, then click **Partition** in the **Time Series** group.

• Click the **Data_PartitionTS1** worksheet, then click **Smoothing – Holt Winters**.

• Click **Additive**.

• Select **Time Variable** and selected variable, then click OK.
Third Session Summary

• Prediction – predict the value of continuous outcome from independent variables.
  • XLMiner prediction techniques.
  • Fitting prediction models to data.
  • Working with output of each method.
  • Applying fitted models to predict response for new observations.
• Time Series Forecasting – predict value of a continuous outcome based on past values in the same series.
  • Smoothing techniques.
• Associate – find relationships between variables.
Every action in business generates data which can be a valuable strategic asset for decision-making.

Data mining enables you to find and extract useful information, discover patterns, and gain insight from your datasets.

The ability to use data intelligently is a vital skill for business analysts.

XLMiner gives you all the tools you need to visualize and transform your data in Excel, and later apply supervised and unsupervised learning methods.

XLMiner is a part of Analytic Solver Platform - a complete toolset for descriptive, predictive and prescriptive analytics.
Contact Info

• Dr. Sima Maleki
• Best way to contact me: Consulting@Solver.com
• You may also download this presentation from our website.
• You can download a free trial version of XLMiner at http://www.solver.com/xlminer-data-mining
References

• MANAGEMENT SCIENCE-The Art of Modeling with Spreadsheets, 4th Edition

• DATA MINING FOR BUSINESS INTELLIGENCE

• Spreadsheet Modeling and Decision Analysis: A Practical Introduction to Business Analytics, 7th Edition
  http://www.cengage.com/us/

• Essentials of Business Analytics, 1st Edition
  http://www.cengage.com/us/
FRONTLINE
solvErs

Thank You!