AGENDA

MACHINE LEARNING

• Background
• Use cases in healthcare, insurance, retail and banking
• Examples:
  • Unsupervised Learning – Principle Component Analysis
  • Supervised Learning – Support Vector Machines
  • Semi-supervised Learning – Deep Learning
  • Supervised Learning – Ensemble Models
• Resources
WHAT IS MACHINE LEARNING?

Wikipedia:
“Machine learning is a scientific discipline that deals with the construction and study of algorithms that can learn from data. Such algorithms operate by building a model based on inputs and using that to make predictions or decisions, rather than following only explicitly programmed instructions.”

SAS:
Machine learning is a branch of artificial intelligence that automates the building of systems that learn from data, identify patterns, and make decisions – with minimal human intervention.
MACHINE LEARNING  GARTNER HYPE CYCLE (ADVANCED ANALYTICS)

Source: Gartner (July 2015)
MACHINE LEARNING

WHY NOW?

- Powerful computing resources have become available
- Data is a commodity: more and different types of data have become available
- Rise of the “data scientist”

https://www.google.com/trends/
WHY WOULD YOU USE MACHINE LEARNING?

- Machine learning is often used in situations where the **predictive accuracy of a model is more important than the interpretability of a model**.

- Common applications of machine learning include:
  - Pattern recognition
  - Anomaly detection
  - Medical diagnosis
  - Document classification

- Machine learning shares many approaches with statistical modeling, data mining, data science, and other related fields.
MACHINE LEARNING  EVERYDAY USE CASES

- Internet search
- Digital adds
- Recommenders
- Image recognition – tagging photos, etc.
- Fraud/risk
- Upselling/cross-selling
- Churn
- Human resource management
EXAMPLES OF MACHINE LEARNING IN HEALTHCARE, BANKING, INSURANCE AND OTHER INDUSTRIES
- About 232,000 people in the US diagnosed with breast cancer in 2015
- About 40,000 will die from breast cancer in 2015

- 60 different drugs approved by FDA
- Tamoxifen
  - Harsh side effects (including increased risk of uterine cancer)
  - Success rate of 80%

80% effective in 100% of patients? Big Data + Compute Resources + Machine Learning 100% effective in 80% of patients (0% effective in rest)

*Big Data, Data Mining, and Machine Learning, Jared Dean*
$10,000 Parkinson’s Data Challenge

6000 hours of data from 16 subjects (9 PD, 7 Control)

Data collected from mobile phones:
- Audio
- Accelerometry
- Compass
- Ambient
- Proximity
- Battery level
- GPS
MACHINE LEARNING IN HEALTHCARE

MEDICATION ADHERENCE

• 32 million Americans use three or more medicines daily

*Drugs don't work in patients who don't take them*
C. Everett Koop, MD

• 75% of adults are non-adherent in one or more ways
• The economic impact of non-adherence is estimated to cost $100 billion annually

AiCure uses mobile technology and facial recognition to determine if the right person is taking a given drug at the right time.

Face recognition
Medication identification
Ingestion

HIPAA-compliant network
Berg, a Boston-area startup, studies healthy tissues to understand the body’s molecular and cellular natural defenses – and what leads to a disease’s pathogenesis.

It’s using concepts of machine learning and big data to scope out potential drug compounds – ones that could have more broad-ranging benefits, pivoting away from today’s trend toward targeted therapies.  

MedCityNews
Telematics driving behavior captured from a mobile app

46 Variables: Speed, acceleration, deceleration, jerk, left turns, right turns, ...

Goal: Stratify observations into low to high risk groups

Cluster analysis and principle component analysis is used
The model assigns a ‘probability score’ of default (PF) to each merchant for a possible fraud risk.

PF score warns the management in advance of probable future losses on merchant accounts.

Banks rank order merchants based on their PF score, and focus on the relatively riskier set of merchants.

The model can capture 62 percent frauds in the first decile.
### Netflix Data Set

<table>
<thead>
<tr>
<th>User</th>
<th>movie 1</th>
<th>movie 2</th>
<th>movie 3</th>
<th>movie 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>User A</td>
<td>3</td>
<td>2</td>
<td>?</td>
<td>5</td>
</tr>
<tr>
<td>User B</td>
<td>?</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>User C</td>
<td>1</td>
<td>?</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>User D</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>User E</td>
<td>1</td>
<td>3</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

- 480K users
- 18K movies
- 100M ratings 1-5 (99% ratings missing)

### Goal:

$1M$ prize for 10% reduction in RMSE over Cinematch

**BelKor’s Pragmatic Chaos**

Declared winners on 9/21/2009

Used ensemble of models, an important ingredient being Low-rank factorization
• Another form of recommender

• Only send to those who need incentive
<table>
<thead>
<tr>
<th>Machine Learning Term</th>
<th>Multidisciplinary Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case, instance, example</td>
<td>Observation, record, row, data point</td>
</tr>
<tr>
<td>Feature, input</td>
<td>Independent variable, variable, column</td>
</tr>
<tr>
<td>Label</td>
<td>Dependent variable, target</td>
</tr>
<tr>
<td>Class</td>
<td>Categorical target variable level</td>
</tr>
<tr>
<td>Train</td>
<td>Fit</td>
</tr>
<tr>
<td>Score</td>
<td>Predict</td>
</tr>
</tbody>
</table>
MACHINE LEARNING BACKGROUND

MULTIDISCIPLINARY NATURE OF BIG DATA ANALYSIS

- Statistics
- Pattern Recognition
- Computational Neuroscience
- Data Science
- Databases
- KDD
- Data Mining
- Machine Learning
- AI
In semi-supervised learning, supervised prediction and classification algorithms are often combined with clustering.
PCA IS USED IN COUNTLESS MACHINE LEARNING APPLICATIONS

- Fraud detection
- Word and character recognition
- Speech recognition
- Email spam detection
- Texture classification
- Face Recognition
Can sea bass and salmon be distinguished based on length?

Length is not a good classifier!
Can sea bass and salmon be distinguished based on lightness?

Lightness provides a better separation.
SEA BASS OR SALMON?

Width and lightness used jointly

Almost perfect separation!
Key questions:

• Should we use all available features in the analysis?

• What happens when there is feature redundancy or correlation?

When the number of features is large, they are often correlated. Hence, there is redundancy in the data.
Principal component analysis (PCA) converts a set of possibly correlated features into a set of linearly uncorrelated features called principal components.

Principal components are:

• Linear combinations of the original variables

• The most meaningful basis to re-express a data set
WHAT IS PRINCIPAL COMPONENT ANALYSIS?

Reduce the dimensionality:
- Reduce memory and disk space needed to store the data
- Reveal hidden, simplified structures
- Solve issues of multicollinearity
- Visualize higher-dimensional data
- Detect outliers
The first principle component is found according to two criteria:

- The variation of values along the principal component direction should be maximal.
- The reconstruction error should be minimal if we reconstruct the original variables.
PRINCIPAL COMPONENT ANALYSIS
FOR MACHINE LEARNING

TWO-DIMENSIONAL EXAMPLE

PCA.mp4
TWO-DIMENSIONAL EXAMPLE

Original Data

PCA Output

PRINCIPAL COMPONENT ANALYSIS FOR MACHINE LEARNING
**PRINCIPAL COMPONENT ANALYSIS FOR MACHINE LEARNING**

**NO DIMENSION REDUCTION: 100 → 100**

Original data: 10,000 observations, 100 variables

Transformed data: 10,000 observations, 100 variables

Projection matrix includes all 100 principal components

\[
10,000 \times 100 \quad \text{Y} \quad = \quad 10,000 \times 100 \quad \text{X} \quad \text{P}
\]
DIMENSION REDUCTION: 100 → 2

Choose the first two principal components that have the highest variance!

\[ \begin{bmatrix} 10,000 \times 2 \\ 10,000 \times 100 \end{bmatrix} = \begin{bmatrix} 100 \times 2 \\ P \end{bmatrix} \]
HOW TO CHOOSE THE NUMBER OF PRINCIPAL COMPONENTS?

PROC PRINCOMP DATA=SASHELP.iris;
  VAR SepalLength SepalWidth PetalLength PetalWidth;
RUN;

96% of the variance in the data is explained by the **first two** principal components.
Training set consists of 100 images... For each image... Face image 50x50

Face vector 2500x1

Calculate average face vector

Subtract average face vector from each face vector

100 face vectors
Training set consists of 100 images

Calculate principal components of $X$, which are eigenvectors of $C_x = X^t X$
Training set consists of 100 images

2500 principle components, each 2500 x 1 dimensional

Select the 18 principle components that have the largest variation

18 selected principle components (eigenfaces)
Principal Component Analysis for Machine Learning

Face Recognition Example

Represent each image in the training set as a linear combination of eigenfaces.

\[
\begin{bmatrix}
w_1 \\
w_2 \\
\vdots \\
w_{18}
\end{bmatrix}
\]  \\
\[
\begin{bmatrix}
w_1 \\
w_2 \\
\vdots \\
w_{18}
\end{bmatrix}
\]

Weight vector for image 1  \\
Weight vector for image 100
1. Convert the unknown image as a face vector
2. Standardize the face vector
3. Represent this face vector as a linear combination of 18 eigenfaces
4. Calculate the distance between this image’s weight vector and the weight vector for each image in the training set
5. If the smallest distance is larger than the threshold distance → unknown person
6. If the smallest distance is less than the threshold distance, then the face is recognized as →
EXAMPLE 2: SUPERVISED LEARNING

SUPPORT VECTOR MACHINES
Construct a hyperplane that maximizes margin between two classes.

Support vectors: points that lie on (define) the margin.

Maximize margin.
Introduce a penalty $C$ based on distance from misclassified points to their side of the margin

Tradeoff:
- **Wide margin** = more training points misclassified but generalizes better to future data
- **Narrow margin** = fits the training points better but might be overfit to the training data
SEPARATION DIFFICULTIES

Many data sets cannot effectively be separated linearly. Use a "kernel function" to map to a higher-dimensional space.

Circle-in-the-Square data set

2 classes: In, Out

http://techlab.bu.edu/classer/data_sets
SUPPORT VECTOR MACHINES

KEY TUNING PARAMETERS

HP SVM node in SAS Enterprise Miner
- Distributes margin optimization operations distributed
- Exploits all CPUs using concurrent threads

Penalty C (default 1)

Linear

Polynomial (2 or 3 degrees)

Radial basis function

Sigmoid function

Kernel and kernel parameters
EXAMPLE 3: SEMI-SUPERVISED LEARNING

NEURAL NETWORKS & DEEP LEARNING
Neural network compute node

$f$ is the so-called activation function

In this example there are four weights $w$’s that need to be determined
The prediction formula for a NN is given by

\[ P(Y \mid X) = g(T_Y) \]
\[ T_Y = \beta_{0Y} + \beta_Y^T Z \]
\[ Z_m = \sigma(\alpha_{0m} + \alpha_m^T X) \]

The functions \( g \) and \( \sigma \) are defined as

\[ g(T_Y) = \frac{e^{T_Y}}{e^{T_N} + e^{T_Y}} \quad , \quad \sigma(x) = \frac{1}{1 + e^{-x}} \]

In case of a binary classifier \( P(N \mid X) = 1 - P(Y \mid X) \)

The model weights \( \alpha \) and \( \beta \) have to be estimated from the data
NEURAL NETWORKS  ESTIMATING THE WEIGHTS

Back propagation algorithm

- Randomly choose small values for all $w_i$’s
- For each data point (observation)
  1. Calculate the neural net prediction
  2. Calculate the error $E$ (for example: $E = (\text{actual} - \text{prediction})^2$)
  3. Adjust weights $w$ according to:

\[
 w_{\text{new}}^{i} = w^{i} + \Delta w^{i}
\]

\[
 \Delta w^{i} = -\alpha \frac{\partial E}{\partial w^{i}}
\]

4. Stop if error $E$ is small enough.
### Supervised Learning Models
- **Single-layer network:**
  - Multilayer perceptron (MLP)
- **Deep Network:**
  - Deep belief network or Deep MLP

### Unsupervised Learning Models
- **Unsupervised Models**
  - **Autoencoder**
    - (with noise injection: denoising autoencoder)
  - **Stacked autoencoder**
    - (with noise injection: stacked denoising autoencoder)
• **Neural networks** with **many layers** and different types of …
  • Activation functions.
  • Network architectures.
  • Sophisticated optimization routines.
• **Each layer represents an optimally weighted, non-linear combination of the inputs.**
• **Extremely accurate** predictions using deep neural networks
• The most common applications of deep learning involve **pattern recognition** in unstructured data, such as **text**, **photos**, **videos** and **sound**.
Unsupervised neural networks that use inputs to predict the inputs

OUTPUT = INPUT

INPUT

ENCODE

x₁  x₂  x₃

h₁₁  h₁₂

x₁  x₂  x₃

DECODE

x₁  x₂  x₃

Note: Linear activation function corresponds with 2 dimensional principle components analysis

Many separate, unsupervised, single hidden-layer networks are used to initialize a larger unsupervised network in a layerwise fashion.
DEEP LEARNING  DIGIT RECOGNITION - CLASSIC MNIST TRAINING DATA

- 784 features form a 28x28 digital grid
- Greyscale features range from 0 to 255
- 60,000 labeled training images  
  (785 variables, including 1 nominal target)
- 10,000 unlabeled test images  
  (784 input variables)
First Two Principal Components

Output of middle hidden layer from 400-300-100-2-100-300-400 Stacked Denoising Autoencoder
DEEP LEARNING | MEDICARE PROVIDERS DATA

Medicare Data:
- Billing information
- Number of severity of procedures billed
- Whether the provider is a university hospital
- …

Goals:
- Which Medicare providers are similar to each other?
- Which Medicare providers are outliers?
Methodology to analyze the data:

• Create 16 k-means clusters
• Train simple denoising autoencoder with 5 hidden layers
• Score training data with trained neural network
• Keep 2-dimensional middle layer as new feature space
• Detect outliers as points from the origin of the 2-dimensional space
DEEP LEARNING

VISUALIZATION AFTER DIMENSION REDUCTION
Investment Portfolio
“Wisdom of the crowd” – Aristotle (‘Politics’)

- collective wisdom of many is likely more accurate than any one
CONCEPT

High variance model (overfit)

High bias model (underfit)

Ensemble (average)
ENSEMBLE MODELING FOR MACHINE LEARNING

CONCEPT

- Combine strengths
- Compensate for weaknesses
- Generalize for future data

Model with Sample #1

Model with Sample #2

Ensemble (average)
Arguably, the Netflix Prize’s most convincing lesson is that a disparity of approaches drawn from a diverse crowd is more effective than a smaller number of more powerful techniques” – Wired magazine

http://www.netflixprize.com/

Predictive modeling competitions - Teams very often join forces at later stages
COMMON USAGE

- Forecasting
  - Election forecasting (polls from various markets and demographics)
  - Business forecasting (market factors)
  - Weather forecasting

http://www.nhc.noaa.gov/
COMMON USAGE

• Meteorology ensemble models for forecasting

http://www.wral.com/fishel-weekend-snow-prediction-is-an-outlier/14396482/
ASSIGNING OUTPUT

- Interval Targets – Average value
- Categorical Targets – Average of posterior probabilities OR majority voting

### Posterior probabilities:
Yes = 0.4, No = 0.6

<table>
<thead>
<tr>
<th>Model</th>
<th>Probability of YES</th>
<th>Probability of NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>Model 2</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>Model 3</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>Model 4</td>
<td>0.35</td>
<td>0.65</td>
</tr>
<tr>
<td>Model 5</td>
<td>0.3</td>
<td>0.7</td>
</tr>
</tbody>
</table>

0.47

0.53

### Spread of ensemble member results gives indication of reliability

https://communities.sas.com/thread/78171
Ways to incorporate multiple models:

- One algorithm, different data samples
- One algorithm, different configuration options
- Different algorithms
- Expert knowledge
APPROACHES

One algorithm, different data samples

**Bagging (Bootstrap Aggregating)** – parallel training and combining of base learners

Average/Vote for final prediction
RANDOM FORESTS

- Ensemble of decision trees
- Score new observations by majority vote or average

Each tree is trained from a sample of the full data.

Variable candidates for splitting rule are random subset of all variables (bias reduction).
One algorithm, different data samples

Boosting

Weight misclassified observations

Each iteration trains a model to predict the residuals of the previous model (i.e., model the error)

Boosting degrades with noisy data – weights may not be appropriate from one run to the next
One algorithm, different configuration options
One algorithm, different configuration options
**ENSEMBLE MODELING FOR MACHINE LEARNING**

**MULTIPLE ALGORITHMS**

Different algorithms

Super Learners
Combine cross-validated results of models from different algorithms
ENSEMBLE MODELING FOR MACHINE LEARNING

IMPROVED FIT STATISTICS

<table>
<thead>
<tr>
<th>Model</th>
<th>Misclassification Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble</td>
<td>0.151</td>
</tr>
<tr>
<td>Bagging</td>
<td>0.185</td>
</tr>
<tr>
<td>Boosting</td>
<td>0.161</td>
</tr>
</tbody>
</table>
An Introduction to Machine Learning
http://blogs.sas.com/content/sascom/2015/08/11/an-introduction-to-machine-learning/

SAS Data Mining Community
https://communities.sas.com/data-mining

SAS University Edition (FREE!)

Products Page

“Overview of Machine Learning with SAS Enterprise Miner”
http://support.sas.com/rnd/papers/sasgf14/313_2014.zip

GitHub
https://github.com/sassoftware/enlighten-apply
https://github.com/sassoftware/enlighten-deep
MACHINE LEARNING  MISCELLANEOUS RESOURCES

- “Big Data, Data Mining, and Machine Learning”

- Cloudera data science study materials
  http://cloudera.com/content/cloudera/en/training/certification/ccp-ds/essentials/prep.html

- Kaggle data mining competitions
  http://www.kaggle.com/

- Python machine learning packages
  OpenCV: http://opencv.org/
  Pandas: http://pandas.pydata.org/
  Theano: http://deeplearning.net/software/theano/

- R machine learning task view
  http://cran.r-project.org/web/views/MachineLearning.html

- Quora: list of data mining and machine learning papers