Beginner’s guide to Machine Learning competitions

EuroPython 2015

Christine Doig
Slides  bit.ly/ep2015-ml-tutorial

Christine Doig
Data Scientist, Continuum Analytics

ch_doig
chdoig
chdoig.github.io
Data Science  
Machine Learning  Supervised learning  Classification  
NLP  Sentiment analysis

Anaconda  
Kaggle  
Competitions  
Dataset

Feature preparation  
Modeling  
Optimization  
Validation
Data Science
Machine Learning
Supervised learning
Classification
NLP
Sentiment analysis
Data Science
I IS DATA SCIENTIST
From the lab to the factory - Data Day Texas

Slides: http://www.slideshare.net/joshwills/production-machine-learninginfrastructure

Video: https://www.youtube.com/watch?v=v-9JycakKjc


http://www.experfy.com/blog/become-data-scientist/
data science

Machine Learning/Stats

Scientific Computing

Analytics

Distributed Systems

Web
data science

**Machine Learning/Stats**
- Data Scientists/Modeler
- PyMC
- StatsModels
- scikit learn
- theano

**Analytics**
- Data/Business Analyst
- pandas
- xlwings
- PostgreSQL
- SQLAlchemy

**Scientific Computing**
- Numba
- SciPy
- NumPy
- xray
- PyTables
- Python
- HDF

**Distributed Systems**
- Spark
- Kafka
- mrjob
- Luigi
- Storm
- Hadoop

**Web**
- Flask
- Scrapy
- Bokeh
- Django
- Jupyter
- RDFLib

**Developer**
Machine Learning
MACHINE LEARNING
Machine Learning

Unsupervised learning
- Clustering
  - K-means
  - Hierarchical clustering
  - *Topic modeling
- Latent variables/structure
  - Dimensionality reduction
  - *Topic modeling

Supervised learning
- Categorical
  - Classification
  - Logistic regression
  - SVM
  - Decision trees
  - k-NN
- Quantitative
  - Regression
  - Linear regression

no labels

labels
Machine Learning

Unsupervised learning
- Clustering
  - K-means
  - Hierarchical clustering
  - *Topic modeling
- Latent variables/structure
  - Dimensionality reduction
  - *Topic modeling

Supervised learning
- Classification
  - Logistic regression
  - SVM
  - Decision trees
  - k-NN
- Regression
  - Linear regression

*Labels include categorical and quantitative data.
Machine Learning

Unsupervised learning

- no labels

- Exploratory
  - group similar individuals together

Supervised learning

- labels

- Predictive
  - Classification
    - predict whether an individual is going to buy/click or not
  - Regression
    - predict how much is the individual going to spend

Unsupervised learning examples:

<table>
<thead>
<tr>
<th>id</th>
<th>gender</th>
<th>age</th>
<th>job_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F</td>
<td>67</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>45</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>F</td>
<td>18</td>
<td>2</td>
</tr>
</tbody>
</table>

Supervised learning examples:

<table>
<thead>
<tr>
<th>id</th>
<th>gender</th>
<th>age</th>
<th>job_id</th>
<th>buy/click_ad</th>
<th>money_spent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F</td>
<td>67</td>
<td>1</td>
<td>Yes</td>
<td>$1,000</td>
</tr>
<tr>
<td>2</td>
<td>M</td>
<td>32</td>
<td>2</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
<td>45</td>
<td>1</td>
<td>No</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>F</td>
<td>18</td>
<td>2</td>
<td>Yes</td>
<td>$300</td>
</tr>
</tbody>
</table>
Natural Language Processing
Machine Learning

Natural language processing

Field concerned with the interactions between computers and human (natural) languages

Sentiment analysis

Extract subjective information on polarity (positive or negative) of a document (text, tweet, voice message…)

e.g. online reviews to determine how people feel about a particular object or topic.
Machine Learning

Unsupervised learning

- Clustering
  - K-means
  - Hierarchical clustering
  - *Topic modeling

- Latent variables/structure
  - Dimensionality reduction
  - *Topic modeling

Supervised learning

- Categorical
  - Classification
    - Logistic regression
    - SVM
    - Decision trees
    - k-NN

- Quantitative
  - Regression
    - Linear regression

Sentiment analysis

- Movie review
  - Positive
  - Negative

E.g.

* denotes tools related to topic modeling.
I love you!  Positive 😊
Setup
Setup

Anaconda

Kaggle

Competitions

Dataset
Setup options

You already have Python installed and your own workflow to install Python packages

Install dependencies in README

Anaconda

Free Python distribution with a bunch of packages for data science

Miniconda

Python + conda (package manager) + packages

http://continuum.io/downloads

git clone git@github.com:chdoig/ep2015-ml-tutorial.git
cd ep2015-ml-tutorial
conda env create
source activate ep-ml
Kaggle

https://www.kaggle.com/

hosts online machine learning competitions
Kaggle Competition

https://www.kaggle.com/c/word2vec-nlp-tutorial

### Bag of Words Meets Bags of Popcorn

Tue 9 Dec 2014 – Tue 30 Jun 2015 (18 days ago)

#### Data Files

<table>
<thead>
<tr>
<th>File Name</th>
<th>Available Formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>sampleSubmission</td>
<td>.csv (276.17 kb)</td>
</tr>
<tr>
<td>unlabeledTrainData.tsv</td>
<td>.zip (25.98 mb)</td>
</tr>
<tr>
<td>testData.tsv</td>
<td>.zip (12.64 mb)</td>
</tr>
<tr>
<td>labeledTrainData.tsv</td>
<td>.zip (12.96 mb)</td>
</tr>
</tbody>
</table>

#### Data Set

The labeled data set consists of 50,000 IMDB movie reviews, specially selected for sentiment analysis. The sentiment of reviews is binary, meaning the IMDB rating < 5 results in a sentiment score of 0, and rating >=7 have a sentiment score of 1. No individual movie has more than 30 reviews. The 25,000 review labeled training set does
Kaggle Competition

Bag of Words Meets Bags of Popcorn

https://www.kaggle.com/c/word2vec-nlp-tutorial

Data

50,000 IMDB movie reviews

- labeledTrainData.tsv
  25,000 rows containing an id, sentiment, and text for each review.
- testData.tsv
  25,000 rows containing an id and text for each review

Task

predict the sentiment for each review in the test data set
Process

Feature preparation

Modeling

Optimization

Validation
**Feature preparation**

**Feature extraction**

the process of making features from available data to be used by the classification algorithms

<table>
<thead>
<tr>
<th>id</th>
<th>sentiment</th>
<th>review</th>
<th>count_words</th>
<th>terrible_word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>the movie was terrible</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>I love it</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Awesome! Love it!</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>I hated every minute</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
Feature extraction

Text

Tokenization

Stopwords

Simple
transition, metal, oxides, considered, generation, materials, field, electronics, advanced, catalysts, tantalum, v, oxide, reports, synthesis, material, nanometer, size, unusual, properties…

Collocations
transition, metal_oxides, considered, generation, materials, field, electronics, advanced, catalysts, tantalum, oxide, reports, synthesis, material, nanometer_size, unusual, properties, sol_gel_method, biomedical_applications…

Entities
transition, metal_oxides, tantalum, oxide, nanometer_size, unusual_properties, dna, easy_method, biomedical_applications

Combination
transition, metal_oxides, generation, tantalum, oxide, nanometer_size, unusual_properties, sol, dna, easy_method, biomedical_applications

Lemmatization
transition, metal, oxide, consider, generation, material, field, electronic, advance, catalyst, property…

language
generic

a
above
across
after
afterwards
again
against
all

domain
specific

material
temperature
advance
size
...

...
Vector Space

Dictionary

1 - transition
2 - metal
3 - oxides
4 - considered
...

Corpus - Bag of words

\[((0, 1), (1, 1), (2, 1))\]
\[((0, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1))\]
\[((2, 1), (5, 1), (7, 1), (8, 1))\]
\[((1, 1), (5, 2), (8, 1))\]
\[((3, 1), (6, 1), (7, 1))\]
\[((9, 1))\]
\[((9, 1), (10, 1))\]
\[((9, 1), (10, 1), (11, 1))\]
\[((4, 1), (10, 1), (11, 1))\]
Naive Bayes Classifier

\[
P(A|B) = P(B|A) \times P(A) / P(B)
\]

<table>
<thead>
<tr>
<th>id</th>
<th>sentiment</th>
<th>review</th>
<th>count_words</th>
<th>terrible_word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>the movie was terrible</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>I love it</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Awesome! Love it</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>I hated every minute</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

What’s the probability of the review being positive if the word love appears in the review?

\[
P(1 \mid \text{love}) = P(\text{love} \mid 1) \times P(1) / P(\text{love}) = \frac{(2/2 \times 2/4)}{(2/4)} = 100%
\]
Modeling.ipynb
Overfitting occurs whenever a model learns from patterns that are present in the training data but do not reflect the data-generating process. Seeing more than is actually there. A kind of data hallucination.

Validation

Hold out method

Training data

Test data

accuracy
Crossvalidation

Accuracy = average(Round1, Round 2, \ldots)

Accuracy in each round with validation set

Training + Validation

Test

Final Accuracy \textit{one shot at this!}
Confusion matrix

Validation

Real

Model prediction

Positive reviews

Accuracy

95%

5%

95%
### Confusion Matrix

<table>
<thead>
<tr>
<th>model/real</th>
<th>positive</th>
<th>negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>95</td>
<td>5</td>
</tr>
<tr>
<td>negative</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
ROC curve/ AUC

100% true positive  
0 % false positive

true positive

false positive
ROC curve/ AUC

100% true positive
0 % false positive

true positive

false positive

AUC
### Public Leaderboard - Bag of Words Meets Bags of Popcorn

This leaderboard is calculated on all of the test data.

<table>
<thead>
<tr>
<th>#</th>
<th>Team Name</th>
<th>Score</th>
<th>Entries</th>
<th>Last Submission UTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Zhang Yang</td>
<td>1.0000</td>
<td>2</td>
<td>Mon, 08 Jun 2015 06:01:54</td>
</tr>
<tr>
<td>2</td>
<td>Puma Gigliotti</td>
<td>1.0000</td>
<td>14</td>
<td>Sat, 20 Jun 2015 02:34:35 (-2.1d)</td>
</tr>
<tr>
<td>3</td>
<td>Abner</td>
<td>1.0000</td>
<td>36</td>
<td>Thu, 25 Jun 2015 07:19:13</td>
</tr>
<tr>
<td>4</td>
<td>Cristian</td>
<td>0.99999</td>
<td>7</td>
<td>Tue, 30 Jun 2015 23:39:07</td>
</tr>
<tr>
<td>5</td>
<td>HeChen</td>
<td>0.99996</td>
<td>1</td>
<td>Tue, 05 May 2015 13:59:05</td>
</tr>
<tr>
<td>6</td>
<td>Alejandro Peláez</td>
<td>0.99259</td>
<td>25</td>
<td>Sun, 10 May 2015 01:10:37 (-0.7h)</td>
</tr>
<tr>
<td>7</td>
<td>Sebastian Raschka_</td>
<td>0.99156</td>
<td>5</td>
<td>Fri, 19 Jun 2015 18:13:32</td>
</tr>
<tr>
<td>8</td>
<td>vgng</td>
<td>0.97663</td>
<td>22</td>
<td>Sun, 28 Jun 2015 23:24:02 (-7.8d)</td>
</tr>
</tbody>
</table>
## Ensemble methods

### Classifier 1  Classifier 2  Classifier 3

<table>
<thead>
<tr>
<th>id</th>
<th>cls_1</th>
<th>cls_2</th>
<th>cls_3</th>
<th>ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Example methods:
- **Majority voting**
- **Weighted voting**
Ensemble.ipynb
Hi all,

Since the competition is over, it would be great to have a topic to discuss the solutions.

My submission has the LB score of 0.97663. It is based on the ensemble of NBSVM, Paragraph Vector and Gated Recurrent Neural Network. The code is at: https://github.com/vinhkhuc/kaggle-sentiment-popcorn.git

I also would like to know about other approaches.
Concepts
- Data Science
- Machine Learning
- Supervised learning
- Classification
- NLP
- Sentiment analysis

Setup
- Anaconda
- Kaggle
- Competitions
- Dataset

Process
- Feature preparation
- Modeling
- Optimization
- Validation
Machine Learning

Feature preparation
- Feature extraction
- Feature selection
- Feature discretization
- Feature scaling
- Feature imputation

Modeling
- Neural Networks
- Logistic Regression
- SVM
- Naive Bayes classifier
- Decision trees

Optimization
- Hyperparameters
- Bagging
- Boosting
- Ensemble
- Random forest
- Regularization
- Hold out method
- Crossvalidation
- Confusion matrix
- ROC curve / AUC

Validation
- Ensemble
- ROC curve / AUC

Machine Learning
Q&A