SEMANTIC COMPUTING

Lecture 5: Introduction to Machine Learning

Dagmar Gromann
International Center For Computational Logic

TU Dresden, 14 May 2018
Overview

- Definition and Applications of Machine Learning
- Supervised Machine Learning
  - Classification and Regression
  - Overfitting and Underfitting
Basic Introduction To Machine Learning
Definition Machine Learning

**Tom M. Mitchell**

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.”

**Alan Turing**

“Can machines think?”

**More general**

A set of paradigms and algorithms that target the detection of regularities in real-world observable records in order to infer patterns and extract insights. The goal in most tasks is to **predict** outcomes for new data (previously unseen by the model).
Machine Learning Applications

- Image Recognition
- Text Classification
- Document Categorization
- Speech Recognition
- Spam Detection
- Fraud Detection
- NLP
- Playing Games
- ...

Dagmar Gromann, 14 May 2018  Semantic Computing  5
Modeling

Model

is a specification of mathematical (or probabilistic) relationships between different variables.

Model: Cooking recipes
Input: “Number of guests” and “quantity of food per guest”
Output: Quantity of ingredients
Some Machine Learning Paradigms

- **Supervised learning**: learning with labeled examples
- **Semi-supervised or weakly supervised learning**: learning with labeled and unlabeled data
- **Unsupervised learning**: learning with unlabeled data
- **Distant supervision**: automatically generate labeled data (by creating examples from existing resources, e.g. large knowledge bases such as Freebase or sentiment lexicons)
- **Reinforcement learning**: learning with indirect or delayed feedback (trial and error / reward and punishment)
Supervised Machine Learning
Data

Data sets

**Training set**: data annotated or usually hand-labeled with the correct category/answer

**Test set**: new examples that test the generalizability of the model trained on the training set
Supervised Learning Process

Training Set → Learning Algorithm → Test Set → h → Prediction

hypothesis
Typical Supervised Machine Learning Problems

- **Classification**: from data to discrete classes
- **Regression**: predicting a continuous value
Classification Example

Demo: Pepper Multi-Modal Emotion Classification and First Date with Pepper

Image source: https://invidis.de/2018/02/digitalisierung-airports-lufthansa-und-airport-muenchen-testen-roboter-pepper-mit-ai/

Dagmar Gromann, 14 May 2018
Regression Example

A diabetes dataset describes 10 physiological variables (see below) measured on 442 patients and an indication of the progression of the disease after one year.

```python
from sklearn import datasets
diabetes = datasets.load_diabetes()
print(diabetes.feature_names)
```

<table>
<thead>
<tr>
<th>age</th>
<th>sex</th>
<th>Body mass index</th>
<th>average blood pressure</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>59</td>
<td>2</td>
<td>32.1</td>
<td>101</td>
<td>252</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>19.2</td>
<td>87</td>
<td>137</td>
</tr>
<tr>
<td>48</td>
<td>1</td>
<td>21.6</td>
<td>87</td>
<td>75</td>
</tr>
</tbody>
</table>

The task is to predict disease progression from physiological variables.

Source original dataset: https://www4.stat.ncsu.edu/~boos/var.select/diabetes.tab.txt

Dagmar Gromann, 14 May 2018
Polynomial Curve Fitting

Fit the data using a polynomial function:

\[ x_1 = \text{age}; x_2 = \text{sex}; x_3 = \text{bmi}; x_4 = \text{bp} \]

\[ m = 441 \text{ patients} \]

\[ h(x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 \]

\[ h(x) = \sum_{i=0}^{n} w_i x_i \]

Values of the coefficient are determined by fitting the polynomial to the training data. This fitting is optimized by minimizing the error function between function \( h(x) \) and any given \( w \).

\[ E(w) = \frac{1}{2} \sum_{i=1}^{m} (h_w(x) - y)^2 \]

Goal: minimize \( E(w) \); a function called sum of the squares
Feature Selection

**Features** are the inputs we provide to our model

Can be provided:

Problem: Predict someone's salary based on the years of experience they have

Which features do we have here?

Answer: only the years of experience
Overfitting and Underfitting

**Overfitting**
Producing a model that performs very well on the data you used for training but generalizes poorly on new data.

**Underfitting**
Producing a model that does not perform well, not even on the training data.
Regression Example

\[ t = \sin(2\pi x) + \text{random noise} \]

Green line = curve of the function \( \sin(2\pi x) \) without noise; red = polynomial function that we fit to the training data


Dagmar Gromann, 14 May 2018
Classification Example

Example dataset for binary classification in sklearn.

```python
from sklearn import datasets
cancer = datasets.load_breast_cancer()
list(data.target_names)

['malignant', 'benign']
```

Example of overfitting/underfitting in a classification setting.
Avoid Overfitting

- **Cross-Validation**: One way to find prediction errors is to use k-fold cross validation where the partition used as a test set varies with each iteration.
- **Early Stopping**: stop the training procedure at the point of the smallest error with respect to the validation set
- **Pruning**: pruning removes nodes from decision trees that add little predictive power to the problem at hand
- **Regularization**: introducing a penalty term to the error function in order to discourage the coefficients from reaching large values
One Way to Avoid Overfitting

k-Fold Cross-Validation:

- Create K training and test sets ("folds") within training set
- For each iteration of K run classifier and test performance against k test set
• What are some important machine learning paradigms?
• What is supervised machine learning?
• For which scenarios could supervised machine learning be used?
• What are important problems of supervised machine learning?
• How can overfitting be avoided and what is it?