Lecture 6: Machine Learning

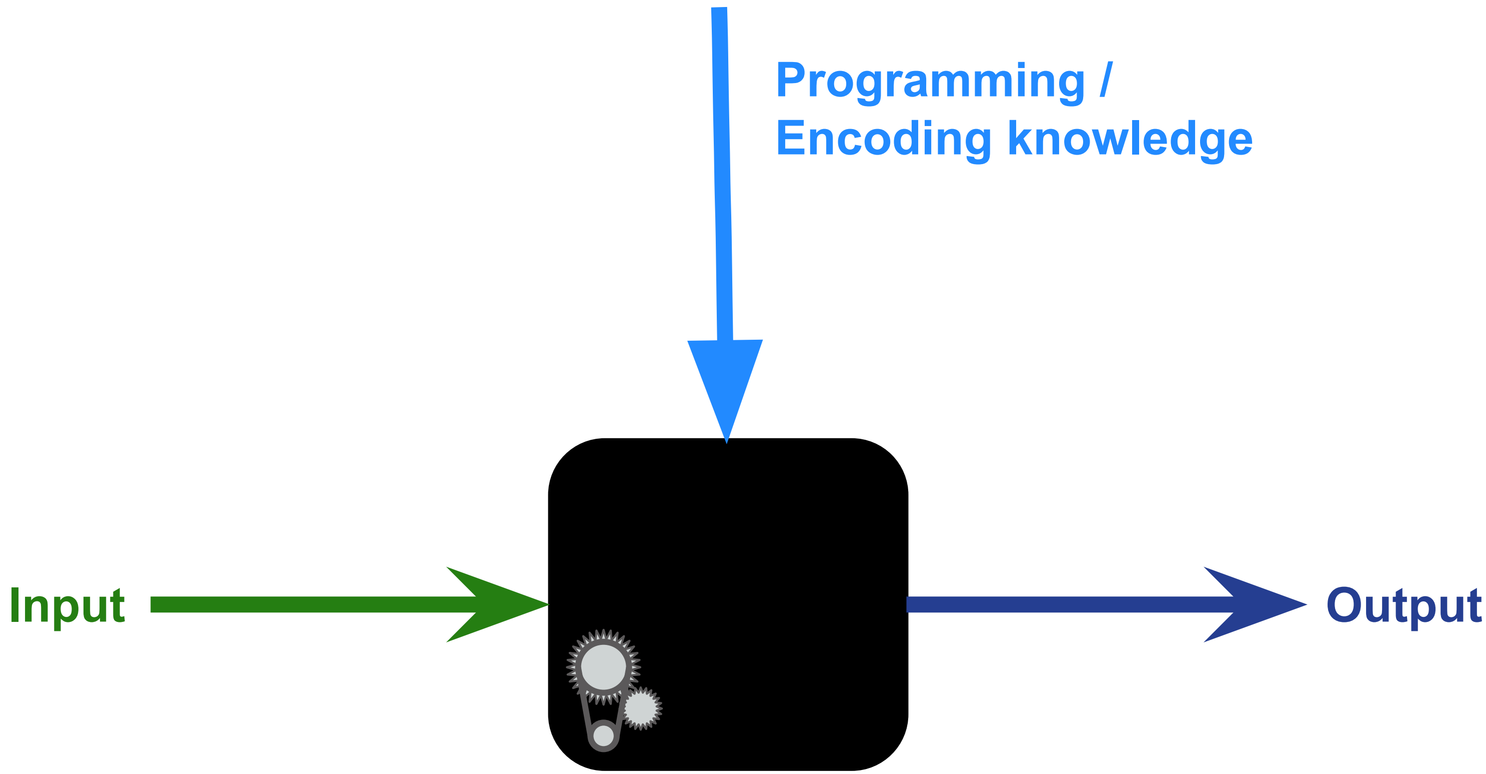
PCL II, CL, UZH
April 6, 2016

Machine Learning

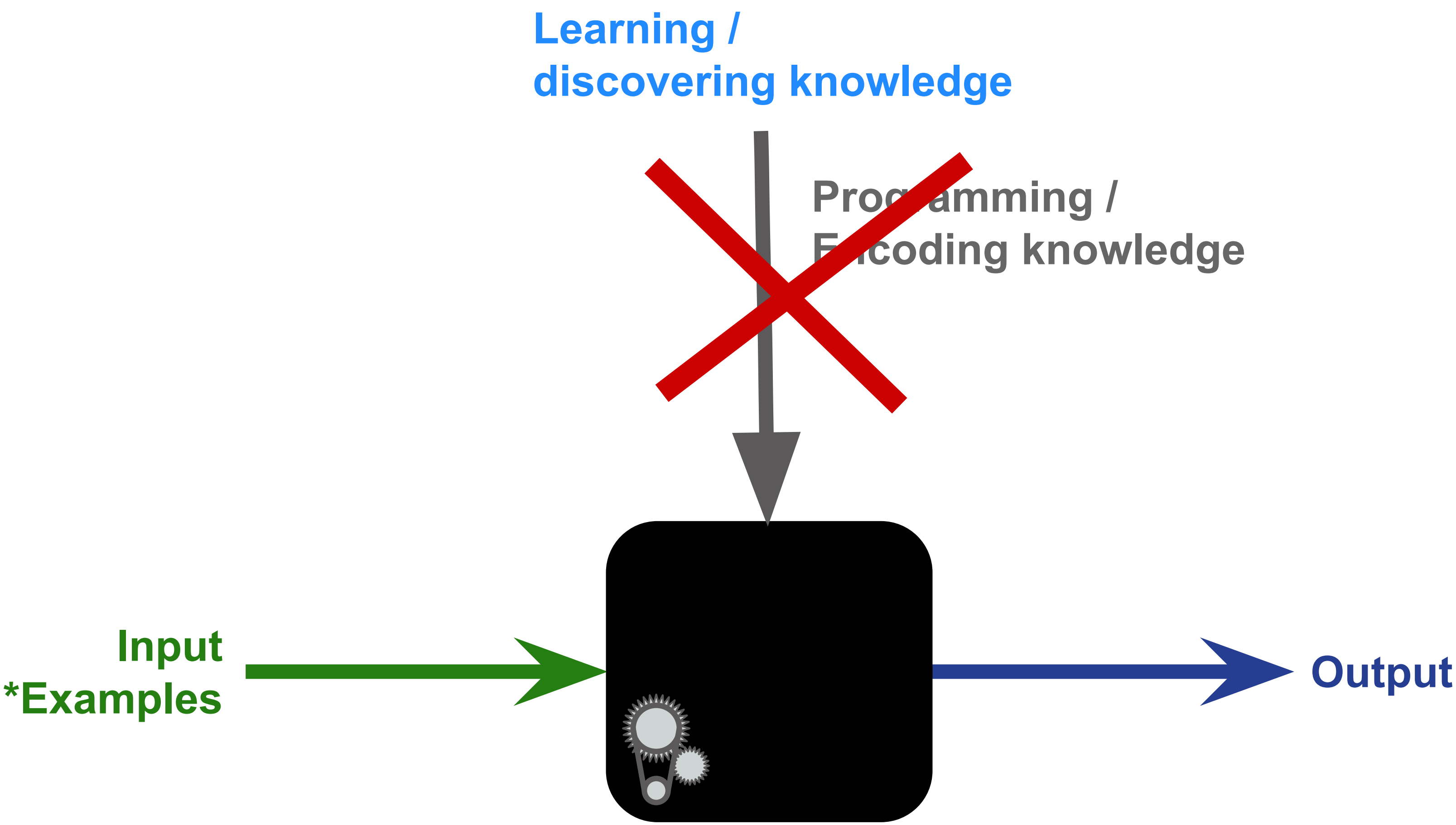
What is Machine Learning?

Machine Learning
Definition

Universität Zürich[UZH]



Machine Learning Definition



Machine Learning

What does “learning from examples” mean?

Contents

Features
Defining Features

Universität Zürich[UZH]

Problem: filter out spam

NO SPAM

From: "Patrick Meyer " <p.meyer@ifi.uzh.ch>

Subject: **Submission of project note**

Yes, let's talk about the overview after these holidays [...]

SPAM

From: ObyhYtatiz@bk.ru

Subject: **Money**

Hi Let me introduce Myself my name Loretta and I know the Secret of making money $4844 if you WAnt to know the Secret follow this [...]

Features
Feature extraction

Problem: filter out spam

NON-SPAM

From: "Patrick Meyer " <p.meyer@ifi.uzh.ch>

Subject: **Submission of project note**

Yes, let's talk about the overview after these holidays [...]

SPAM

From: ObyhYtatiz@bk.ru

Subject: **Money**

Hi Let me introduce Myself my name Loretta and I know the Secret of making money $4844 if you WAnt to know the Secret follow this [...]

F_1	F_2	...	F_N	Class (or label)
V_1	V_2	...	V_N	NON-SPAM
V_1	V_2	...	V_N	SPAM
[...]				

[...]

Features Types

- Numeric
  - Discrete
  - Continuous
- Nominal
  - binary (2 values)
  - N values
- Hybrid

Feature engineering. This is the most important/difficult part

Contents

Universität Zürich[UZH]

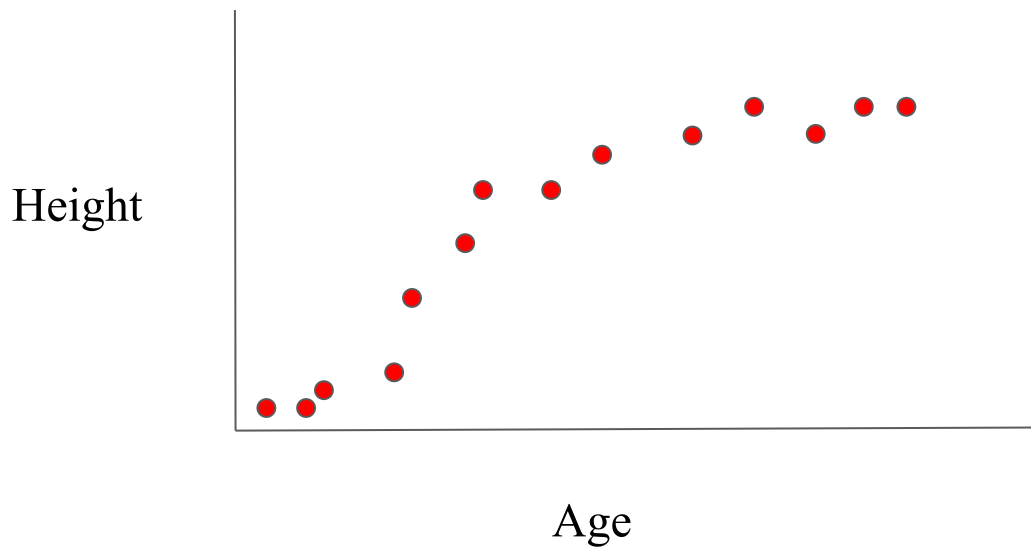
Machine Learning Types

Universität Zürich[UZH]

- Supervised learning - Labeled training data
- Unsupervised learning - Unlabeled training data

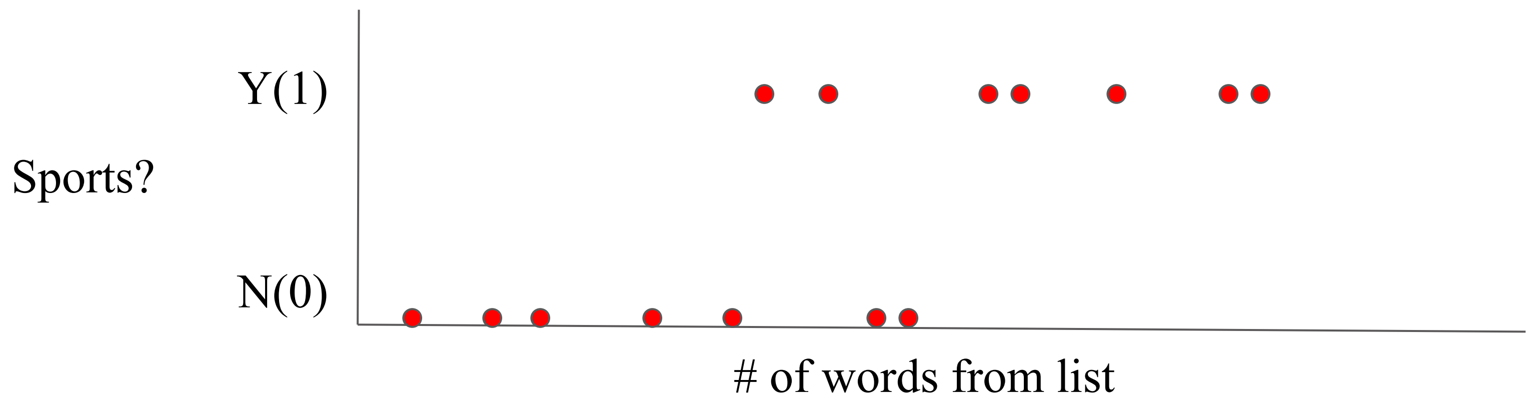
Machine Learning Types
Supervised Learning

Universität Zürich[UZH]



Regression - continuous valued output

Machine Learning Types
Supervised Learning



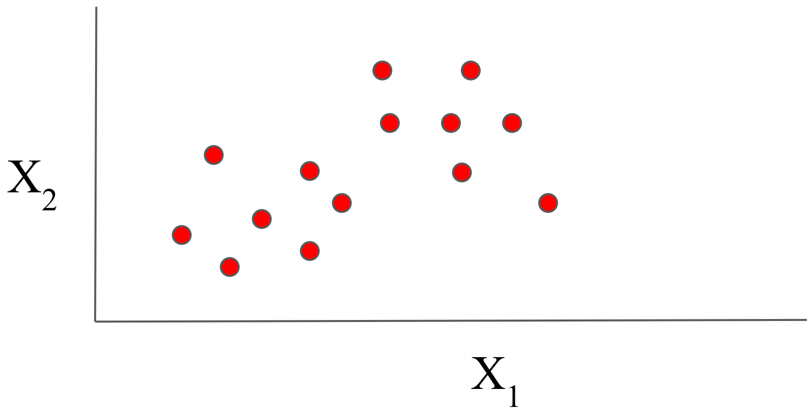
Classification - discrete valued output

Machine Learning Types
Supervised Learning

- Regression - predicts continuous valued output
- Classification - predicts discrete valued output
  - Spam filtering (spam / non-spam)
  - OCR (A / B / a / 0 / 1 / % / …)
  - Word sense disambiguation ($\text{bank}_{\text{river\_edge}}$/$\text{bank}_{\text{institution}}$/…)
  - Document Classification (news/technical/law/…)

Machine Learning Types
Unsupervised Learning

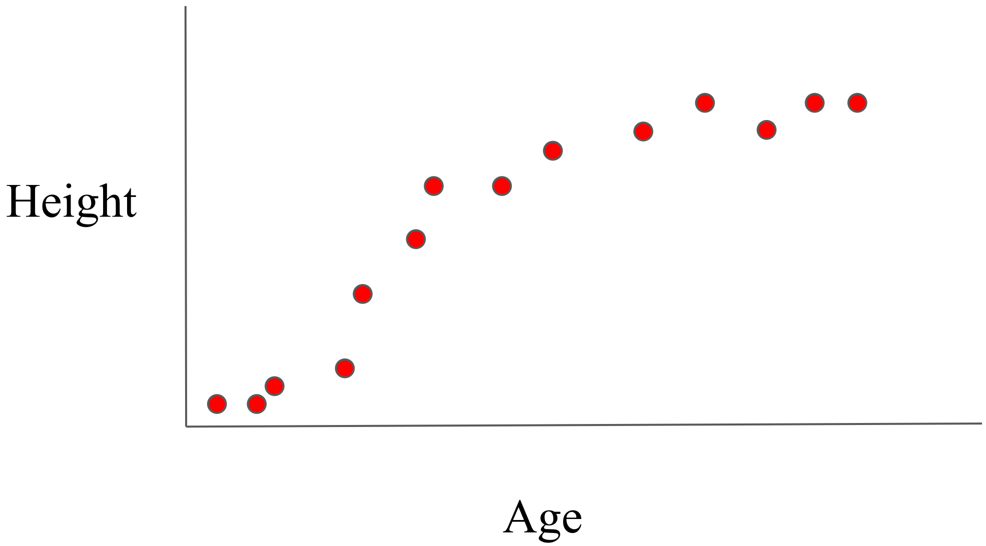
- Unsupervised learning - Unlabeled training data
  - Clustering - K-means algorithm



Contents

Learning Algorithms

Learning Algorithms look for the hypothesis that generalizes best over a training set.



Learning Algorithms
Naive Bayes Classifier

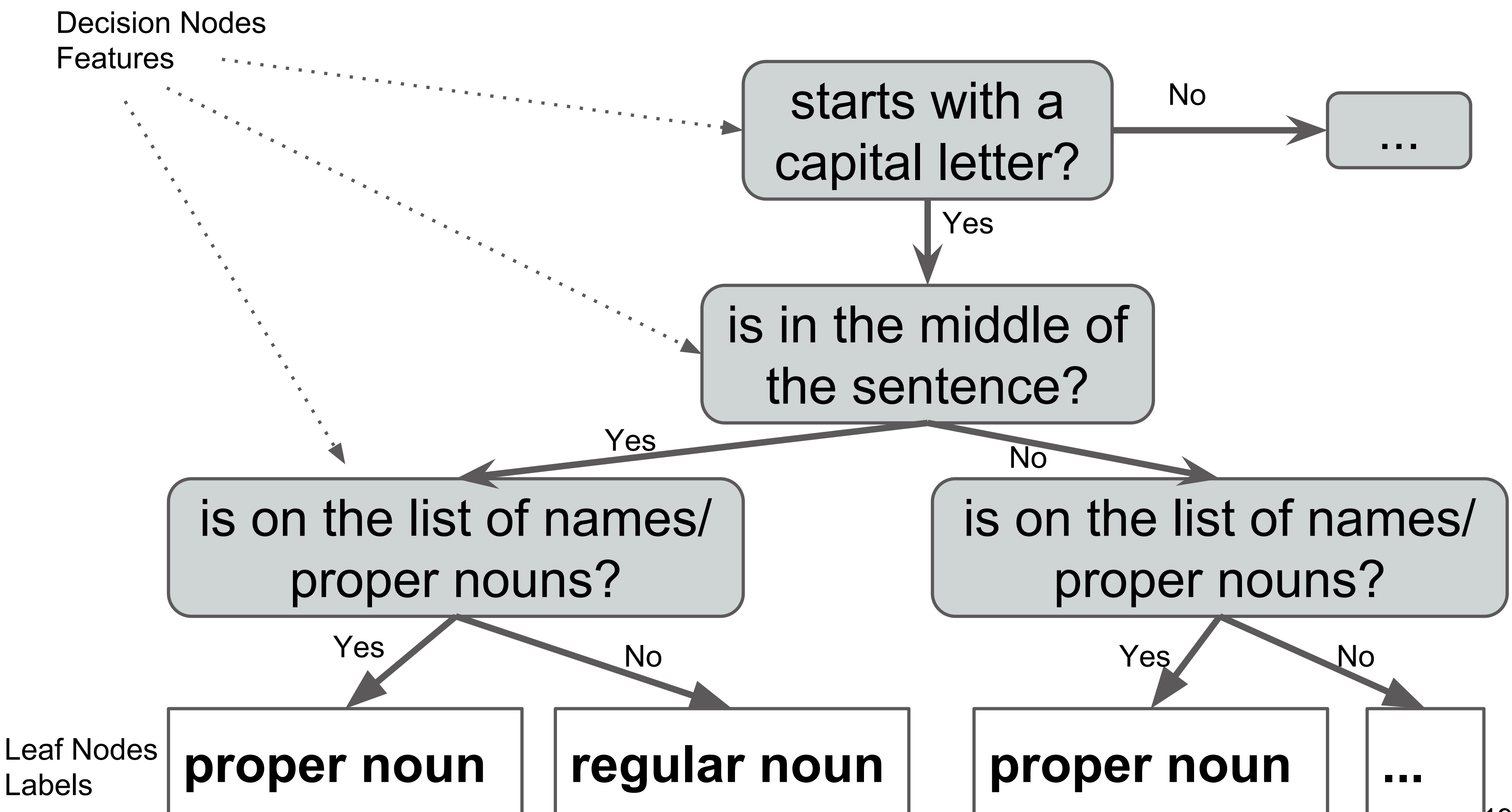
Universität Zürich[UZH]

- It looks for the label that maximizes P(T | *features*)
  1. Compute prior probability to each label p(T')
  2. Combine the contribution of each feature with the prior probability:

     $$T = \mathrm{argmax}_{T'}\, p(T') \times p(f_1 \mid T') \times p(f_2 \mid T') \times \ldots$$

- Learning search: analytical → MLE

Learning Algorithms
Decision Trees



Learning Algorithms
Decision Trees

- Learning: select feature that reduces total entropy the most
  - Entropy of a set:
    $H = -\sum_{output} p(output) \cdot \log p(output)$
  - Little variety – entropy is low
  - Mixed output values – entropy is high
- Learning type: iterative

Learning Algorithms
Decision Trees - Entropy

Universität Zürich[UZH]

```
import math
def entropy(labels):
    freqdist = nltk.FreqDist(labels)
    probs = [freqdist.freq(l) for l in nltk.FreqDist(labels)]
    return -sum([p * math.log(p,2) for p in probs])

>>> print entropy(['male', 'male', 'male', 'male'])
0.0
>>> print entropy(['male', 'female', 'male', 'male'])
0.811278124459
>>> print entropy(['female', 'male', 'female', 'male'])
1.0
>>> print entropy(['female', 'female', 'male', 'female'])
0.811278124459
>>> print entropy(['female', 'female', 'female', 'female'])
0.0
```

Contents

Evaluation
Split Data

Universität Zürich[UZH]

Data splitting

- Training data set ≈ 80%
- Development (or tuning) data set ≈ 10%
- Testing (or evaluation) data set ≈ 10%

Cross validation

- Divide corpus into *N* subsets (folds)
  - 10-fold cross validation

Evaluation

Universität Zürich[UZH]

Accuracy

- Percentage of things right.
- Be careful with unbalanced corpus

Precison and Recall

System	Actual target	Actual ¬target
selected	*tp*	*fp*
¬selected	*fn*	*tn*

precision = tp/(tp+fp)   recall = tp/(tp+fn)

Evaluation
Confusion Matrix

Universität Zürich[UZH]

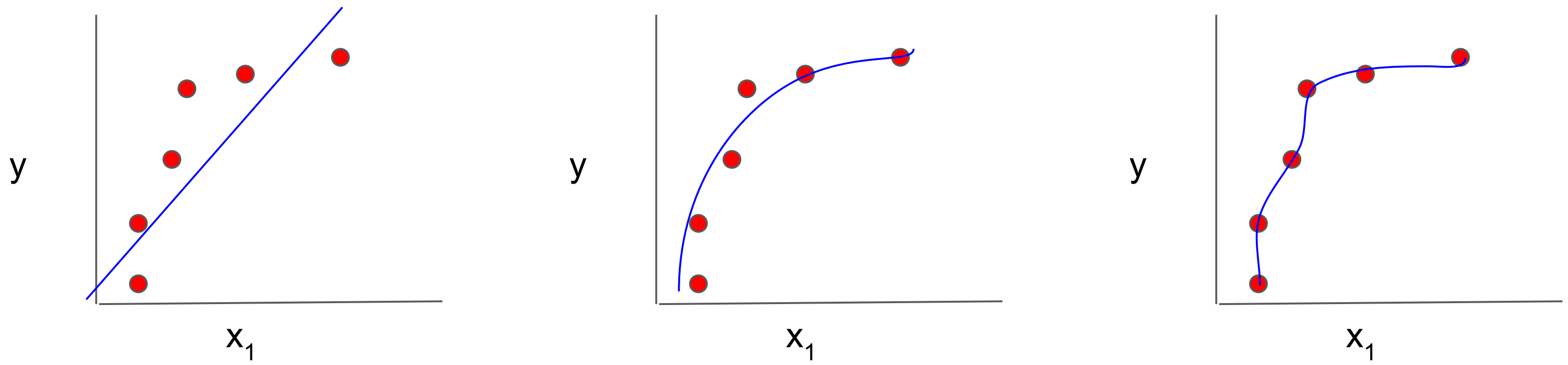
Each cell [*i*,*j*] indicates how often label *j* was predicted when the correct label was *i*

	NN	IN	AT	JJ	.	NNS	,	VB	NP
NN	<11.8%>	0.0%	.	0.2%	.	0.0%	.	0.3%	0.0%
IN	0.0%	<9.0%>	.	.	.	0.0%	.	.	.
AT	.	.	<8.6%>	.	.	.	.	.	.
JJ	1.7%	.	.	<3.9%>	.	.	.	0.0%	0.0%
.	.	.	.	.	<4.8%>	.	.	.	.
NNS	1.5%	.	.	.	.	<3.2%>	.	.	0.0%
,	.	.	.	.	.	.	<4.4%>	.	.
VB	0.9%	.	.	0.0%	.	.	.	<2.4%>	.
NP	1.0%	.	.	0.0%	.	.	.	.	<1.8%>

(row = reference; col = test)

[NLTK book, chapter 6]

Evaluation
Possible Problems



Contents

ML in Python
Name genders

Universität Zürich[UZH]

```
#data set
from nltk.corpus import names
import random
raw_male_names = names.words('male.txt') # ['Aamir', 'Aaron',
raw_female_names = names.words('female.txt') # ['Abagael', 'Abagail',
labeled_male_names = [(name, 'male') for name in raw_male_names]
labeled_female_names = [(name, 'female') for name in raw_female_names]

name_set = labeled_male_names + labeled_female_names
random.shuffle(name_set)
#[('Nissy', 'female'), ('Warren', 'male'), ('Olivie', 'female'),...
```

ML in Python
Name genders

Universität Zürich[UZH]

```
def generate_features(name):
"""function to generate features from an input name"""
    return {'last_letter': name[-1], 'len': len(name) }

#extracting features
feature_set = [(gender_features(n), g) for (n,g) in name_set]

#training set, test set
train_set, test_set = feature_set[500:], feature_set[:500]

#train the classifier
classifier = nltk.NaiveBayesClassifier.train(train_set)


#apply the classifier
classifier.classify(gender_features('Neo'))
#'male'

classifier.classify(gender_features('Trinity'))
#'female'

#classifier accuracy on the test set
print nltk.classify.accuracy(classifier, test_set)   #0.758
```

Contents

Universität Zürich[UZH]

Practicalities

1. Get data - Usually the more the better

2. What features can be useful for learning?
   a. If too many features - Feature Selection

3. Choose an algorithm
   a. Labeled or unlabeled data
   b. Regression or Classification
   c. [...]

4. 80% Training, 10% Evaluation, 10% Testing
   a. if data set is large - Just pick Ev. and Test data randomly
   b. Otherwise, 10-fold cross validation

References

Universität Zürich[UZH]

➔ In Python
- `nltk.NaiveBayesClassifier`
  - NLTK book, sections 6.1, 6.5
- `nltk.classify.decisiontree.DecisionTreeClassifier`
  - NLTK book, section 6.4
- `nltk.classify.maxent.MaxentClassifier`
  - NLTK book, chapter 6

➔ Learning tool (algorithms not specially efficient)
- Weka http://www.cs.waikato.ac.nz/ml/weka/

Lecture 6:
Machine Learning

PCL II, CL, UZH
April 6, 2016

Universität Zürich[UZH]