# *Lecture 13:*

Isabelle Guyon
guyoni@inf.ethz.ch

# *Exercise Class Poster*

http://clopinet.com/isabelle/Projects/ETH/homework13.zip

# *Next semester*

---

## Reading group on CAUSALITY

Isabelle Guyon

André Elisseeff

This class will discuss research papers on causality inference from observational or experimental data. The selected papers aim at understanding machine learning techniques to infer causality, including causal graphs derived from "graphical models".

Earn (easily) 4 credit points, for 1 hour of reading group

Tuesday CAB H52, 17h-18h

http://www.vorlesungsverzeichnis.ethz.ch/Vorlesungsverzeichnis/LerneinheitDetailsPre.do?semkez=2006S&leId=33662

# *Homework 13*

- **Baseline methods and poster.**
  1) Download the package for homework 13 (includes software and template poster).
  2) Modify the poster to include you own results.
  3) Make an entry fo Arcene, Dexter, Gisette, and Madelon to the website http://www.nipsfsc.ecs.soton.ac.uk/
  4) Email the text of the poster to guyoni@inf.ethz.ch with subject "Homework13" no later than:
  Tuesday February 7th.

# *Good Posters:*

1) **Good material**
   - know what you want to talk about
2) **Good poster**
   - use a good template
3) **Good communication skills**
   - practice, practice, practice
   - bring business cards and handouts

# Baseline Methods for the Feature Extraction Class

**Isabelle Guyon**

## SUMMARY

The course on feature extraction of the winter semester 2005-2006 covered material from the book "Feature Extraction, Fundamentals and Applications", I. Guyon et al Eds., to appear in Springer. The book presents the results of a challenge organized for the NIPS2003 conference. The students learned about the techniques employed by the best challengers and tried to match or outperform the performances of the best entries. A Matlab® toolbox was provided to them with some sample code. We present the results of the baseline methods of the sample code.

## BACKGROUND

Feature extraction is an essential pre-processing step to pattern recognition and machine learning problems. It is often decomposed into feature construction and feature selection. Classical algorithms of feature construction were reviewed in class. More attention has been given to the feature selection step because of the recent success of methods involving a large number of "low-level" features (image pixels, text "bag-of-word", molecular structural features, gene expression coefficients.)

The course encouraged students to learn practical data analysis techniques to match the results of the best entrants of the NIPS 2003 feature selection challenge. The students could submit their results to the website of the challenge, which is still available for post-challenge entries: http://www.nipsfsc.ecs.soton.ac.uk/.

## DATASETS

The NIPS 2003 challenge in feature selection was to find feature selection algorithms that significantly outperform methods using all features, using as benchmark ALL five datasets provided. The tasks are two-class classification problems. The datasets were split into training, validation, and test set. The participants had initially only access to the labels of the training set. They could obtain immediate feed-back on their validation set performance by submitting their prediction labels on-line. The validation set labels were made available shortly before the end of the challenge. The test labels are still not released to the public.

| Dataset | Size | Type | Features | Training Examples | Validation Examples | Test Examples |
|---|---|---|---|---|---|---|
| Arcene | 8.7 MB | Dense | 10000 | 100 | 100 | 700 |
| Gisette | 22.5 MB | Dense | 5000 | 6000 | 1000 | 6500 |
| Dexter | 0.9 MB | Sparse integer | 20000 | 300 | 300 | 2000 |
| Dorothea | 4.7 MB | Sparse binary | 100000 | 800 | 350 | 800 |
| Madelon | 2.9 MB | Dense | 500 | 2000 | 600 | 1800 |

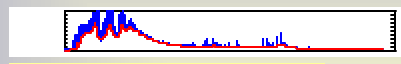
## METHODS

**Scoring:**
The challenge participants were classified according to their balanced error rate (BER) on the test set. The BER is the average of the error rate on the positive class and on the negative class. If two entries had performance not significantly different, the score privileged the entry with the smallest feature set.

**Learning objects:**
Matlab® learning objects are based on two simple abstractions: data and algorithm. The learning object package CLOP based on the Spider developed at the Max Planck Institute can be downloaded from: http://www.modelselect.inf.ethz.ch/models.php.

**Task of the students:**
The students were given baseline methods as CLOP learning objects with given hyper-parameters. The baseline performance for each dataset was about in the tenth percentile of the submissions. The students were asked to get a better BER or fewer features than the baseline method. Extra credit was given for outperforming the best challenge entry. The students were free to use the validation set labels for training.

## RESULTS (BER in %)



### ARCENE: cancer diagnosis


ARCENE Best BER= 11.9±1.2 % - Target featnum=1100 (11%) – Baseline BER=14.7%

```
my_svc=svc({'coef0=1', 'degree=3', 'gamma=0', 'shrinkage=0.1'});
my_model=chain({standardize, s2n('f_max=1100'), normalize, my_svc})
```

TIP#1: train on both validation and test set (BER=22.66% if onlytraining set used).
TIP#2: use ensemble methods (BER=).

### DEXTER: text categorization

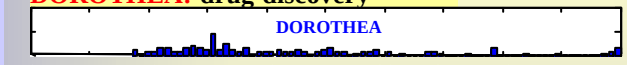
NEW YORK, October 2, 2001 – Instinet Group Incorporated (Nasdaq: INET), the world's largest electronic agency securities broker, today announced that it has completed the acquisition of ProTrader Group, LP, a provider of advanced trading technologies and electronic brokerage services primarily for retail active traders and hedge funds. The acquisition excludes ProTrader's proprietary trading business. ProTrader's 2000 annual revenues exceeded $83 million.


DEXTER Best BER=3.30±0.40% - Target featnum=300 (1.5%) – Baseline BER=5%

```
my_classif=svc({'coef0=1', 'degree=1', 'gamma=0', 'shrinkage=0.5'});
my_model=chain({s2n('f_max=300'), normalize, my_classif})
```

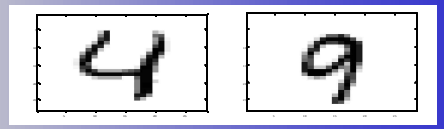
TIP#1: train on both validation and test set (BER=3.95%).
TIP#2: vary the number of features f_max=??? (BER=3.20%).

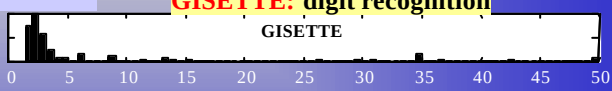
### DOROTHEA: drug discovery




DOROTHEA Best BER=8.54±0.99% - Target featnum=50000 (50%) – Baseline BER=15%

### GISETTE: digit recognition



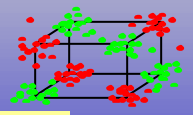



GISETTE Best BER=1.26±0.14% - Target featnum=1000 (20%) – Baseline BER=1.80%

```
my_classif=svc({'coef0=1', 'degree=3', 'gamma=0', 'shrinkage=1'});
my_model=chain({normalize, s2n('f_max=1000'), my_classif});
```

TIP#1: swap s2n and normalize to get better results (BER=1.17%-> 1.11% w. valid set).
TIP#2: use the pixel representation and smooth the data (BER=0.91%).

```
my_classif=svc({'coef0=1', 'degree=4', 'gamma=0', 'shrinkage=0.1'});
my_model=chain({convolve(exp_ker({'dim1=9', 'dim2=9'})), normalize,
my_classif})
```



### MADELON: artificial data


MADELON Best BER=6.22±0.57% - Target featnum=20 (4%) – Baseline BER=7.33%

```
my_classif=svc({'coef0=1', 'degree=0', 'gamma=1', 'shrinkage=1'});
my_model=chain({probe(relief,{'p_num=2000', 'pval_max=0'}), standardize, my_classif})
```

TIP#0: to get the baseline result train with train+valid sets (otherwise get BER=7.89%)
TIP#1: replace pval_max=0 by f_max=20 (the number of features may vary because of the limited precision of the probe method and cause variance in the BER).
TIP#2: vary the number of features fmax=??? (BER=6.67%).

## CONCLUSIONS

The performances of the challengers could be matched or outperformed with the CLOP library without too much difficulty. Simple filter methods (S2N and Relief) were sufficient to get a space dimensionality reduction comparable to what the winners obtained. SVMs are easy to use and generally work better than other methods. We experienced with Gisette to add prior knowledge about the task and could outperform the winners. Further work includes using prior knowledge for other datasets.

# *Proposal*

- **Description:** What is this about?
- **Motivation:** Why should we care?
- **Merit:** Are you the best?
- **Impact:** How is this going to make money or change the world?

# *Report (paper, poster)*

- **Background:** What is this about? Why should we care?
- **Material and methods:** What did we use? How did we proceed?
- **Results:** What did we find out?
- **Conclusion:** Did we succeed or fail? How is this going to change the world?