Lecture 14: Wrap up

Isabelle Guyon
guyoni@inf.ethz.ch

Outline

- Out to the world!
- Lessons from the challenge
- Other types of data
- New algorithms

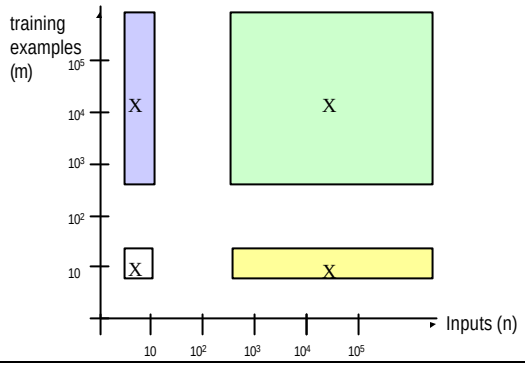
So you have…

- Written a successful proposal
- Just received data
- Now what?
- Sit at your computer and relax!
- Take a look at your data…

Data Statistics

Dataset	X Type	Sparse	Y Type (c)	Balance	Feat	Pat	Pat/Feat
Arcene	cont.	0.459	binary (2)	0.88	10000	100	0.01
Dexter	cont.	0.995	binary (2)	1	20000	300	0.015
Dorothea	binary	0.991	binary (2)	0.195	100000	800	0.008
Gisette	cont.	0.87	binary (2)	1	5000	6000	1.2
Madelon	cont.	0	binary (2)	1	500	2000	4

Data Aspect Ratio



What to do?

- Visualize data:
  - Heat map
  - PCA + scatter plot 2 dim
- Try simple techniques first:
  - One data split (training + validation)
  - Feature ranking with
    - Test/Fisher/s2n (continous)
    - MI/odds ratio (binary) PR (unbalanced binary)
    - Plot pvalue or FDR
    - Scatter plot top ranking features
  - Naïve Bayes classification
  - Linear SVM/ridge regression, with shrinkage ~10^{-5}

What NOT to do?

- Start with a big experiment with
  - many methods
  - systematic hyperparameter search
- Use cross-validation without reserving extra test data.
- Select features with all the data or preprocess all the data, then assess with CV.

Shallow data

Case n>>m

- For best for prediction performance:
  - Prune features with univariate method.
  - Use regularized classification method (i.e. shrinkage/weight decay/ARD prior).
  - Select hyperparameters by CV or use ensemble methods.
- To get a compact set of feature
  - Further prune with embedded methods.

Skinny data

Case m>>n

- Less risk of overfitting, use:
  - Non-linear methods.
  - Wrappers.
  - Bagging.

Big data

Both m and n large

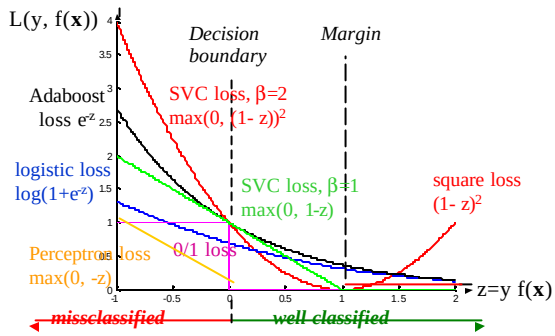
- Prune features with univariate method
- Go to the skinny data case
- Or, use directly embedded methods

My new algorithm

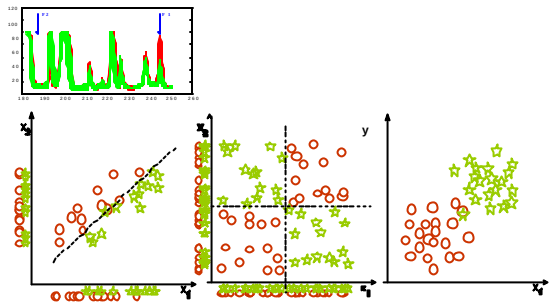
- A new filter
  - Browse through statistics books for great test statistics
  - Invent anything + use the probe method or permutation test
- A new wrapper
  - A new smart search technique
  - A clever assessment method
- A new embedded method
  - Feature scaling + sensitivity or OBD

Picture Gallery by Lectures

1. Loss functions

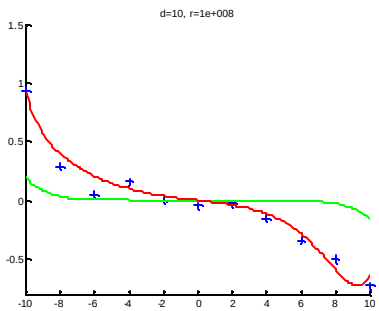


2. Multivariate cases

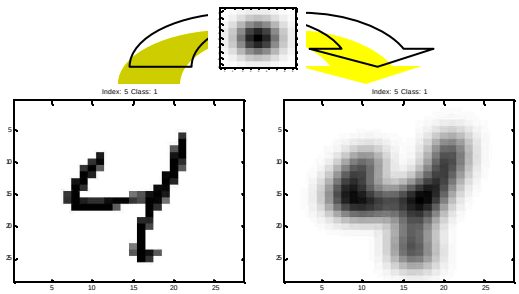


(a) (b) (c)

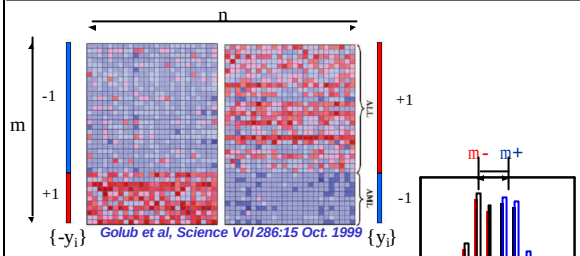
3. Weight decay



4. Convolution



5. S2N



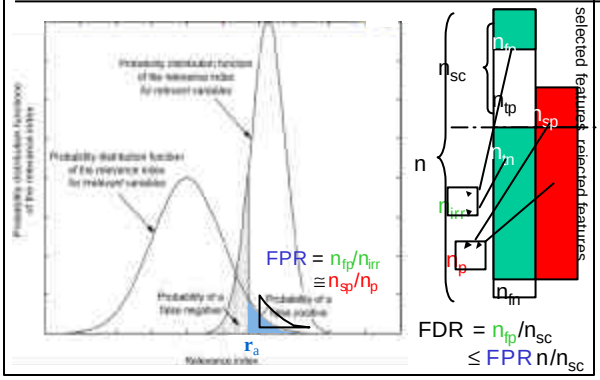
Golub et al, Science Vol 286:15 Oct. 1999

$$S2N = \frac{|\mu_+ - \mu_-|}{\sigma_+ + \sigma_-}$$

$S2N \cong R \sim \mathbf{x} \bullet \mathbf{y}$

after "standardization" $\mathbf{x} \leftarrow (\mathbf{x}-\mu_x)/\sigma_x$

6. Feature significance



FDR = n_{fp}/n_{sc}

$\leq$ FPR n/n_{sc}

7. Kernel "Trick"

- $f(\mathbf{x}) = \mathbf{w} \bullet F(\mathbf{x})$
- $\mathbf{w} = \sum_i \alpha_i y_i F(\mathbf{x}_i)$

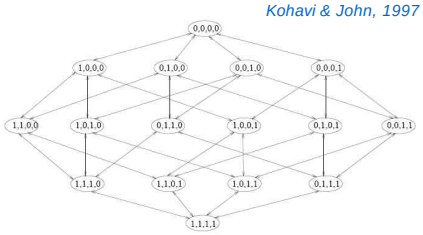
Dual forms

Aizerman-Braverman-Rozonoer -1964

- $f(\mathbf{x}) = \sum_i \alpha_i y_i k(\mathbf{x}_i, \mathbf{x})$
- $k(\mathbf{x}_i, \mathbf{x}) = F(\mathbf{x}_i) \bullet F(\mathbf{x})$

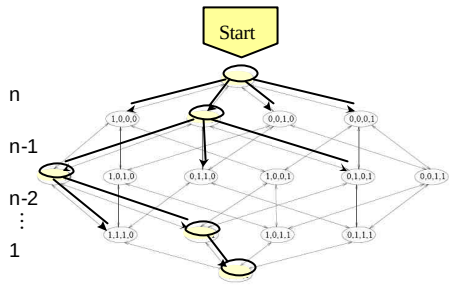
8. Wrappers

Kohavi & John, 1997



Exhaustive search: 2^n trainings, complexity n

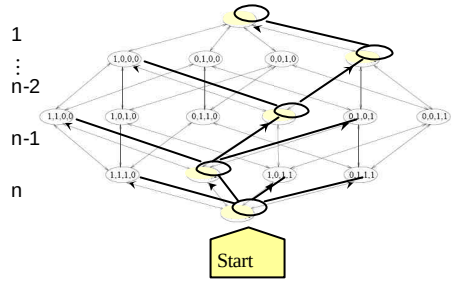
8. Wrappers



Forward selection: n(n+1)/2 trainings, complexity log n

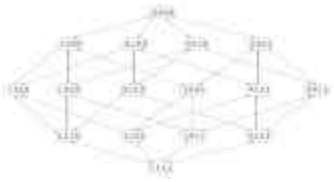
8. Wrappers

Backward elimination: n(n+1)/2 train., complexity log n



9. Embedded Methods

Idea: Transform a discrete space into a continuous space.

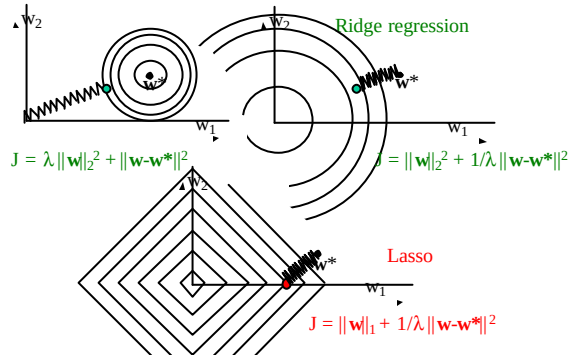


$s=[\sigma_1, \sigma_2, \sigma_3, \sigma_4]$

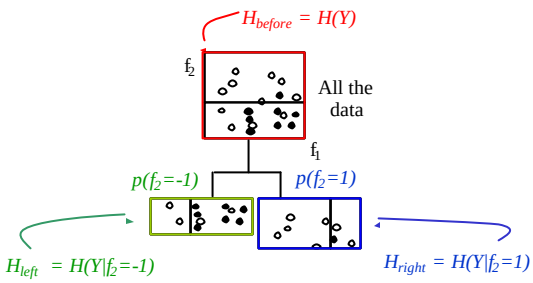
- Discrete indicators of feature presence: $\sigma_i \in \{0, 1\}$
- Continuous scaling factors: $\sigma_i \in \mathbb{R}$

Now we can do gradient descent!

10. Regularizers



11. MI=Information Gain

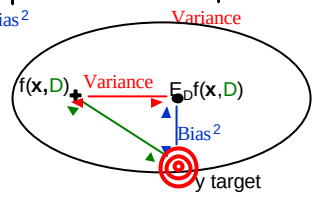


$$\mathbf{IG} = H_{before} - (p(f_2=-1)\, H_{left} + p(f_2=1)\, H_{right}) = H(Y) - H(Y|\, f_2) = \mathbf{MI}$$

12. Bias-variance tradeoff

- D = one training set of size m (m fixed)
- For the square loss:

$$\underbrace{E_D[f(\mathbf{x},D)-y]^2}_{\text{Expected value of the loss over datasets D of the same size}} = \underbrace{[E_D f(\mathbf{x},D)-y]^2}_{\text{Bias}^2} + \underbrace{E_D[f(\mathbf{x},D)-E_D f(\mathbf{x},D)]^2}_{\text{Variance}}$$



13. MAP @ RRM

- Maximum A Posteriori (MAP):

  f = argmax P(f|D)

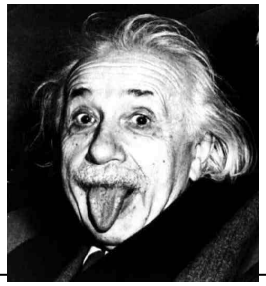
  = argmax P(D|f) P(f)

  = argmin $\underbrace{-\log P(D|f)}_{\text{Negative log likelihood = Empirical risk R[f]}}$ $\underbrace{-\log P(f)}_{\text{Negative log prior = Regularizer } \Omega[f]}$

- Regularized Risk Minimization (RRM):

  f = argmin R[f] + Ω[f]

14. Take your work seriously…



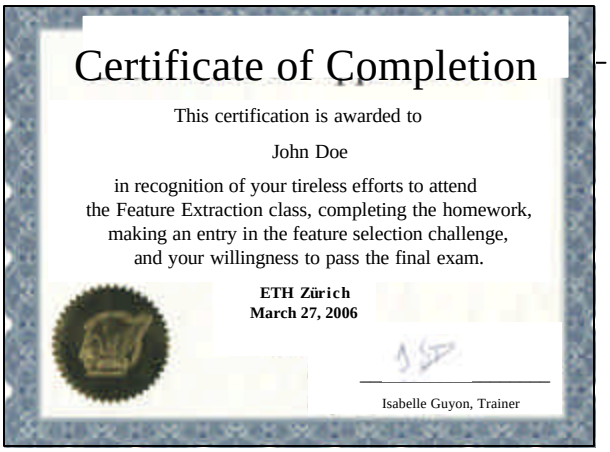
… Don't take yourself seriously!

Exercise Class

Homework 14

- **Make an entry for Dorothea.**
  1) Download the latest version of CLOP and the package for homework 14.
  2) Try to outperform baseline results.
  3) Make an entry for all datasets to the website http://www.nipsfsc.ecs.soton.ac.uk/
  4) Email the URL of the result and the zip file to guyoni@inf.ethz.ch with subject "Homework14" no later than:
  Tuesday February 28th.

Exam Preparation



Where and When?

- **Monday, March 27**
- **CAB G69.2**
- **Register for a time**

What?

1. Challenge submissions (10 points):
   For each dataset, a baseline model is provided having a baseline performance BER0 and number of features n0.
   - **Earn 1 point per dataset for a valid submission having {BER<BER0 , any n} or {BER<=BER0 and n<n0}**
   - **Earn 1 more point per dataset for a submission outperforming the best challenge entry.**
2. Paper presentation (5 points).
3. Final exam, poster + questions (17 points)
   (contents= 4; presentation= 4; questions: 9)

Total = 32 points

Grade = min(6, num_points/4);
Pass: Grade >= 4

How?

- Make a **complete challenge submission** to: http://www.nipsfsc.ecs.soton.ac.uk/
  (all results on all datasets in a single zip).
- Include **descriptions**.
- Email the URL of your exam entry to guyoni@inf.ethz.ch before March 20, 2006

Example:
http://www.nipsfsc.ecs.soton.ac.uk/description/?id=1942

Baseline Results

	BER0 (%)	n0	feat (%)	BestBER±δBER
Arcene	14.7	1100	11	11.9±1.2
Dexter	5	300	1.5	3.30±0.40
Dorothea	15	50000	50	8.54±0.99
Gisette	1.8	1000	20	1.26±0.14
Madelon	7.33	20	4	6.22±0.57

•1 point if {BER<BER0 , any n}
or {BER<=BER0 and n<n0}

•1 more point {BER<BestBER+δBER}

Arcene

Best BER= 11.9 ±1.2 %, n0=1100 (11%), BER0=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#0: train on both validation and test set (BER=22.66% if only training set used).

TIP#1: use ensemble methods.

Dexter

Best BER=3.30±0.40%, n0=300 (1.5%), BER0=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

Best BER=8.54±0.99%, n0=1000 (1%), BER0=12.37%

```
my_model=chain({TP('f_max=1000'), naive,
bias});
```

TIP: try to keep more features with TP and chain with another feature selection method to get overall fewer features.

Gisette

Best BER=1.26±0.14%, n0=1000 (20%), BER0=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%).

Madelon

Best BER=6.22±0.57%, n0=20 (4%), BER0=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})
```

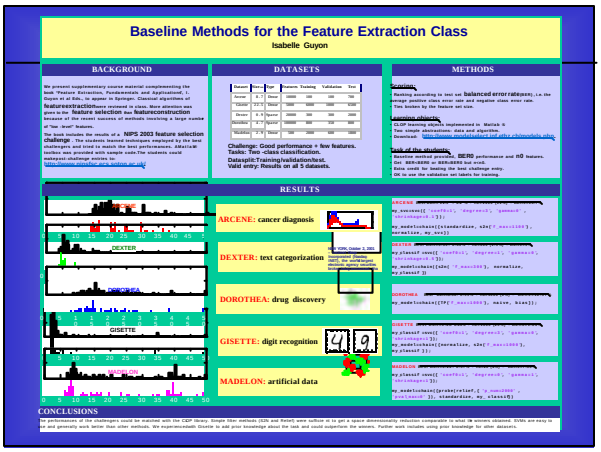
TIP#1: replace pval_max=0 by f_max=20 (num. features vary because of variance of probe method).

TIP#2: vary the number of features fmax=??? (BER=6.67%).

Poster

- **Background:** What is this about? Why should we care?
- **Material:** What did you use?
- **Methods:** How did you proceed?
- **Results:** What did you find out?
- **Conclusion:** Did it succeed or fail? How could we improve?

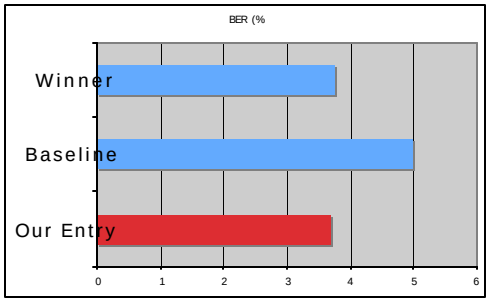
USE PICTURES and GRAPHS



Make Graphs

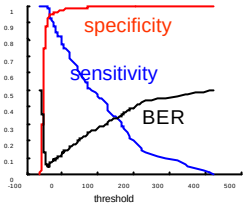
BAR Graphs

- From Theodor (DEXTER)



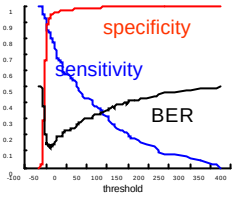
BER = f(threshold)

DOROTHEA



Training set

Theta = -37.40



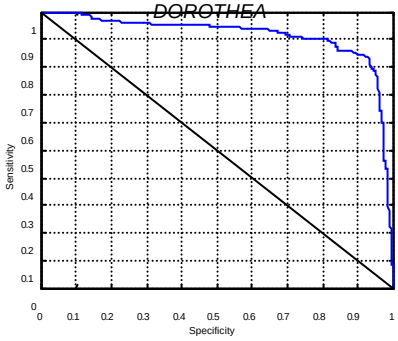
Test set

Theta = -38.14

No bias adjustment, test BER=22.54%; with bias, test BER=12.37%

ROC curve

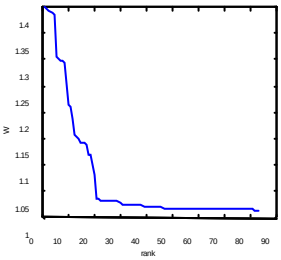
DOROTHEA

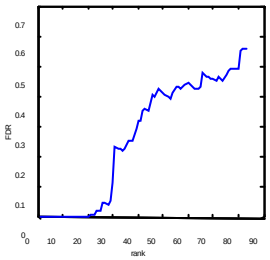


AUC=0.91

Feature Selection

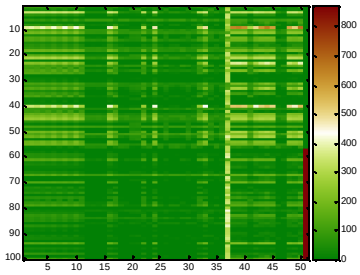
MADELON (pval_max=0.1)





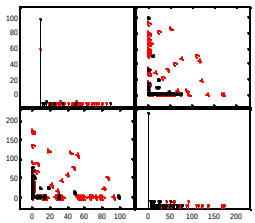
Heat map

ARCENE



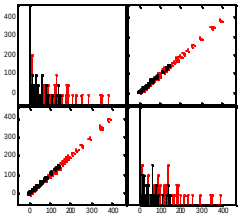
Scatter plots

ARCENE



chain({standardize, s2n('f_max=1100'), normalize, gs('f_max=2'), my_svc})

Test BER=29.37%

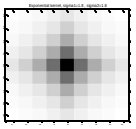


chain({standardize, s2n('f_max=2'), normalize, my_svc})

Test BER=49%

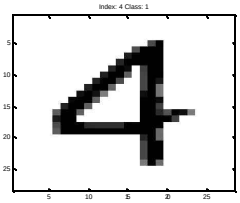
Preprocessing

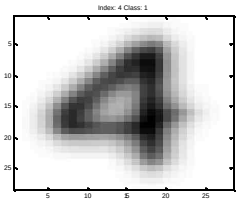
GISETTE (pixelGisette_exp_conv)



```
prepro=my_model{1};
show(prepro.child);
```

```
DD=test(prepro,D.train);
browse_digit(DD.X, D.train.Y);
```





Questions

- Nine questions will be drawn at random among the set of questions found at:

http://clopinet.com/isabelle/Projects/ETH/Exam_Questions.html

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