Lecture 10:

Isabelle Guyon
guyoni@inf.ethz.ch

Maximum Likelihood

- **Likelihood**: *probability of the data given the model*.

  $P(D \mid f) = P(\{(\mathbf{x}_i, y_i)\} \mid f)$

- **Maximum Likelihood (ML)**: find the model that fits best the data.

ML = ERM

- **ML**: $f = \text{argmax } P(D \mid f)$

  $= \text{argmin } -\log P(D \mid f)$

  Negative log likelihood

  $-\log P(D \mid f) = -\log P(\{(\mathbf{x}_i, y_i)\} \mid f)$

  $= \Sigma_i - \log P((\mathbf{x}_i, y_i) \mid f)$

  $= \Sigma_i L(f(\mathbf{x}_i), y_i)$ loss function

  $= R[f]$ Empirical risk

- **ERM**: $f = \text{argmin } R[f]$

Example: Logistic Loss

- Functional margin:

  $z = y_i f(\mathbf{x}_i)$

- Link function:

  $P((\mathbf{x}_i, y_i) \mid f) = \psi(z)$

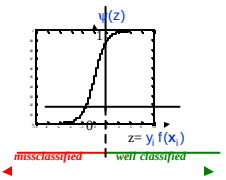
- Logistic link:

  $\psi(z) = 1/ (1+e^{-z})$

- Logistic loss:

  $L(f(\mathbf{x}_i), y_i) = -\log P((\mathbf{x}_i, y_i) \mid f)$

  $= \log (1+e^{-z})$



Priors and Bayesian Learning

- Double random process:
  - Draw a target function f in a family of functions {f}
  - Draw the data pairs ($\mathbf{x}_i$, y_i=f($\mathbf{x}_i$)+noise)
- The distribution of f is called the "prior" P(f).
- Our revised opinion about f once we see the data is the "posterior" P(f|D).
- Bayesian "learning":

  $P(y|x,D) \;\alpha\; \int P(y|x,D,f)\, dP(f|D)$
- MAP:

  $f = \text{argmax } P(f|D)$

  $= \text{argmax } P(D|f)\, P(f)$

MAP = RRM

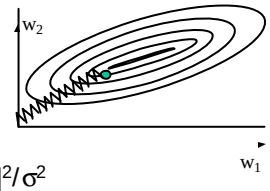
- Maximum A Posteriori (MAP):

  $f = \text{argmax } P(D|f)\, P(f)$

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

  $f = \text{argmin } R[f] + \Omega[f]$

Example: Gaussian Prior



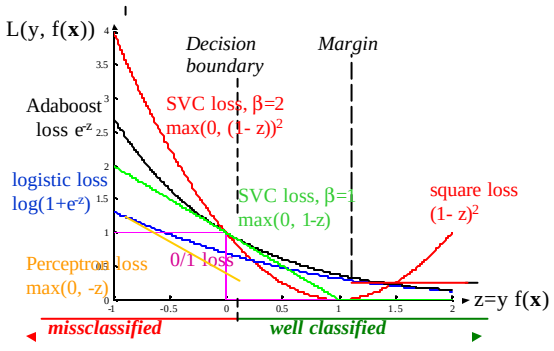
- Linear model:

  $f(\mathbf{x}) = \mathbf{w}.\mathbf{x}$
- Gaussian prior:

  $P(f) = \exp -||\mathbf{w}||^2/\sigma^2$
- Regularizer:

  $\Omega[f] = -\log P(f) = \lambda\, ||\mathbf{w}||^2$

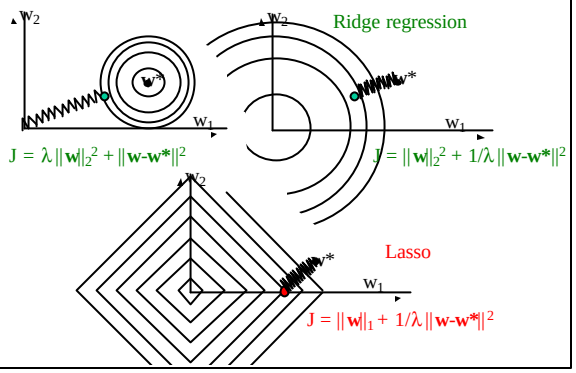
Loss Functions



Regularizers

- $\| \mathbf{w} \|_2^2 = \Sigma_i w_i^2$: 2-norm regularization (ridge regression, original SVM)
- $\| \mathbf{w} \|_1 = \Sigma_i |w_i|$: 1-norm regularization (Lasso Tibshirani 1996, 1-norm SVM 1965)
- $\| \mathbf{w} \|_0 = \text{length}(\mathbf{w})$: 0-norm (Weston et al., 2003)

Mechanical interpretation



Exercise Class
Intellectual Property

http://www.copyright.iupui.edu/IPPrimer.pdf
http://www.ssiplaw.com/publications/patentprimer.pdf

Intellectual Property

- **Why** should you care?
- **What** can you protect?
- **How** can you protect your IP?

Why should you care?

- Want to retain the right of exploiting your invention:
  - prevent others from exploiting it or earn royalties/licensing fees
  - make sure others don't prevent you from exploiting your invention.
- Want to get credit.
- IP rights aim to **promote future innovation** by allowing you to **recoup your investments** in the creative process by providing you a **limited monopoly**.

What can you protect?

- Inventions
- Books/writings/images
- Software
- Works of art
- Brand names, logos

How can you protect your IP?

- Trade secrets
- Publications
- Watermarks
- Copyrights
- Trademarks
- Patents

How to copyright?

- **Copyrightable:**
  - Original (not copied and creative)
  - Fixed (on a material support)
  - Non-functional
  - **Examples**: books, articles, plays, movies, video or sound recordings, art, e-mail messages, computer programs, video games, architectural design
  - **Not**: ideas, facts, processes, discoveries.
- Automatic protection
- Copyright notice (Copyright © Isabelle Guyon, 2005)
- Registration

How to trademark?

- A trademark helps identify the source of a product to prevent consumer confusion.
- **Trademarkable**:
  - Brand names
  - Symbols/logos
  - Slogans
  - Packaging/color/look-and-feel
- Register your trademark
- Use the symbol ® (™ does not confer rights)

How to patent?

- Keep **good records**:
  - Use lab books
  - Document your work with dated reports
  - Keep old software revisions and backups
  - Witness inventions (signature on lab book, digital signature, certified mail)
- Make **no public disclosure** before filing (US 1 year grace period)
- File provisional patent application (US)
- File patent application
  - Switzerland http://www.ige.ch/
  - United States http://www.uspto.gov/

What to patent?

- **Patentable**:
  - "Anything under the sun that is made by man." (Chief Justice Burger of the United States Supreme Court)
    - Apparatus (machine)
    - Method (of preparing a compound, doing business, performing surgery, analyzing data, etc.)
    - Manufactured product or compound.
  - Not: laws of nature, abstract ideas, and physical phenomena, pure mathematic equations.
- Novel
- Useful
- Non-obvious

Structure of a patent

1) Abstract
2) Specification
   - Must enable others skilled in the art to make or utilize the invention.
   - Must include one preferred embodiment.
3) Claims (which conclude the specification)
4) Drawings

Exercise: example of claims

- We want to patent the algorithm of A. Elisseeff from the embedded method exercise class.
- Write a first claim (as broad as possible)
- Write derived claims organized as a tree.

Method to patent (A. Elisseeff)

- Consider the 1 nearest neighbor algorithm. We define the following score:

$$J = \sum_{k=1}^{m} \|x_k - x_{s(k)}\|^2 - \lambda \|x_k - x_{d(k)}\|^2$$

- Where $s(k)$ (resp. $d(k)$) is the index of the nearest neighbor of x_k belonging to the same class (resp. different class) as x_k.

Scaling factors

- $x_{ki} \neg s_i x_{ki}$
- $J = \sum_{k=1}^{m} \sum_{i=1}^{n} s_i^2 [(x_{ki} - x_{s(k)i})^2 - l\,(x_{ki} - x_{d(k)i})^2]$
- $\partial J / \partial s_i = 2\, s_i \underbrace{\sum_{k=1}^{m} [(x_{ki} - x_{s(k)i})^2 - l\,(x_{ki} - x_{d(k)i})^2]}_{Relief\ (i)}$

Uses of the gradient

- $\partial J / \partial s_i$ a s_i *Relief (i)*
- Gradient descent / multiplicative updates:

$s_i \neg s_i - \eta\ s_i$ *Relief (i)* $= s_i\,(1 - \eta)$ *Relief (i)*

- Feature ranking
- Backward elimination
- Forward selection

Claims

1) **Method of analyzing data** (define data):
   - Preprocessing to get features
   - Find nearest hit/miss
   - Compute Relief index
   - Use the index
- **Other claims:**
   - Which preprocessing
   - Which metric
   - Variants of the index
   - Variants of uses (gradient/MU, ranking, forward selection, backward elimination)
   - Visualization
   - Applications

Claim 1

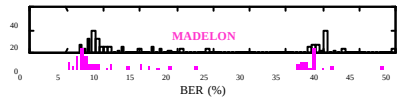
A method of analyzing data, which consist of entities pertaining to a number of categories, the method comprising the steps of:

1. Preprocessing the data to represent the entities as a number of features, such features defining a feature space
2. Endowing the feature space with a metric
3. Identifying for each entity its nearest hit (closest entity of the same category) and its nearest miss (closest entity of another category) according to the metric
4. Evaluating for each entity the discrepancy between the distance to its nearest hit and the distance to its nearest miss, in projection one particular feature
5. Averaging the results of step 4 over all entities for each feature
6. Using the results of step 5 to assess the prevalence of the features.

This week homework…

- Write claims for an algorithm that you have proposed or implemented in one of the previous homework.
- Make an entry for the Madelon dataset using the Relief filter.

Baseline Madelon



Best challenge result: 6.22 ±0.57 % (training on all data)

Baseline method: 7.33 ± 0.61%

```
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})
```

Earn 1 point with a result better than the baseline method.

Earn 2 points with a result better than testBER=6.79%.

Relief vs. Ttest