NeuroComp
Machine Learning and Validation

Michèle Sebag
http://tao.lri.fr/tiki-index.php

Nov. 16th 2011
Validation, the questions

1. What is the result?

2. My results look good. Are they?

3. Does my system outperform yours?

4. How to set up my system?
Contents

Position of the problem
  Background notations
  Difficulties
  The learning process
  The villain

Validation
  Performance indicators
  Estimating an indicator
  Testing a hypothesis
  Comparing hypotheses

Validation Campaign
  The point of parameter setting
  Racing
  Expected Global Improvement
Contents

Position of the problem
  Background notations
  Difficulties
  The learning process
  The villain

Validation
  Performance indicators
  Estimating an indicator
  Testing a hypothesis
  Comparing hypotheses

Validation Campaign
  The point of parameter setting
  Racing
  Expected Global Improvement
Supervised Machine Learning

**Context**

\[ \text{World} \rightarrow \text{instance } x_i \rightarrow \text{Oracle} \downarrow \]
\[ y_i \]

**Input:** Training set \( \mathcal{E} = \{(x_i, y_i), i = 1 \ldots n, x_i \in \mathcal{X}, y_i \in \mathcal{Y}\} \)

**Output:** Hypothesis \( h : \mathcal{X} \mapsto \mathcal{Y} \)

**Criterion:** few mistakes (details later)
Definitions

Example

▶ row : example/ case
▶ column : feature/variables/attribute
▶ attribute : class/label

Instance space \( \mathcal{X} \)

▶ Propositionnal : \( \mathcal{X} \equiv \mathbb{R}^d \)
▶ Relational : ex. chemistry.

<table>
<thead>
<tr>
<th>age</th>
<th>employed</th>
<th>education</th>
<th>edun</th>
<th>marital</th>
<th>...</th>
<th>job</th>
<th>relation</th>
<th>race</th>
<th>gender</th>
<th>hour</th>
<th>country</th>
<th>wealth</th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>State_gov</td>
<td>Bachelors</td>
<td>13</td>
<td>Never_married</td>
<td>...</td>
<td>Admin_clerk</td>
<td>Not_in_farm</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>51</td>
<td>Self_emp</td>
<td>Bachelors</td>
<td>13</td>
<td>Married</td>
<td>...</td>
<td>Exec_mani</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>13</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>39</td>
<td>Private</td>
<td>HS_grad</td>
<td>9</td>
<td>Divorced</td>
<td>...</td>
<td>Handlers</td>
<td>c Not_in_farm</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>54</td>
<td>Private</td>
<td>11th</td>
<td>7</td>
<td>Married</td>
<td>...</td>
<td>Handlers</td>
<td>Husband</td>
<td>Black</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>28</td>
<td>Private</td>
<td>Bachelors</td>
<td>13</td>
<td>Married</td>
<td>...</td>
<td>Prof_speical</td>
<td>Wife</td>
<td>Black</td>
<td>Female</td>
<td>40</td>
<td>Cuba</td>
<td>poor</td>
</tr>
<tr>
<td>58</td>
<td>Private</td>
<td>Masters</td>
<td>14</td>
<td>Married</td>
<td>...</td>
<td>Exec_mani</td>
<td>Husband</td>
<td>White</td>
<td>Female</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>50</td>
<td>Private</td>
<td>9th</td>
<td>5</td>
<td>Married</td>
<td>...</td>
<td>Other_services</td>
<td>Not_in_farm</td>
<td>Black</td>
<td>Female</td>
<td>18</td>
<td>Jamaica</td>
<td>poor</td>
</tr>
<tr>
<td>52</td>
<td>Self_emp</td>
<td>HS_grad</td>
<td>9</td>
<td>Married</td>
<td>...</td>
<td>Exec_mani</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>45</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>31</td>
<td>Private</td>
<td>Masters</td>
<td>14</td>
<td>Married</td>
<td>...</td>
<td>Prof_speical</td>
<td>Not_in_farm</td>
<td>White</td>
<td>Female</td>
<td>50</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>42</td>
<td>Private</td>
<td>Bachelors</td>
<td>13</td>
<td>Married</td>
<td>...</td>
<td>Exec_mani</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>37</td>
<td>Private</td>
<td>Some_col</td>
<td>10</td>
<td>Married</td>
<td>...</td>
<td>Exec_mani</td>
<td>Husband</td>
<td>Black</td>
<td>Male</td>
<td>80</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>30</td>
<td>State_gov</td>
<td>Bachelors</td>
<td>13</td>
<td>Married</td>
<td>...</td>
<td>Prof_speical</td>
<td>Husband</td>
<td>Asian</td>
<td>Male</td>
<td>40</td>
<td>India</td>
<td>rich</td>
</tr>
<tr>
<td>24</td>
<td>Private</td>
<td>Bachelors</td>
<td>13</td>
<td>Married</td>
<td>...</td>
<td>Prof_speical</td>
<td>Own_child</td>
<td>White</td>
<td>Female</td>
<td>30</td>
<td>United_States</td>
<td>rich</td>
</tr>
<tr>
<td>33</td>
<td>Private</td>
<td>Assoc_acv</td>
<td>12</td>
<td>Never_married</td>
<td>...</td>
<td>Sales</td>
<td>Not_in_farm</td>
<td>Black</td>
<td>Male</td>
<td>50</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>41</td>
<td>Private</td>
<td>Assoc_acv</td>
<td>11</td>
<td>Married</td>
<td>...</td>
<td>Craft_repair</td>
<td>Husband</td>
<td>Asian</td>
<td>Male</td>
<td>40</td>
<td>Missing</td>
<td>rich</td>
</tr>
<tr>
<td>34</td>
<td>Private</td>
<td>7th_8th</td>
<td>4</td>
<td>Married</td>
<td>...</td>
<td>Transport</td>
<td>Husband</td>
<td>Amer_Indian</td>
<td>Male</td>
<td>45</td>
<td>Mexico</td>
<td>poor</td>
</tr>
<tr>
<td>26</td>
<td>Self_emp</td>
<td>HS_grad</td>
<td>9</td>
<td>Never_married</td>
<td>...</td>
<td>Farming</td>
<td>Husband</td>
<td>Amer_Indian</td>
<td>Male</td>
<td>35</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>33</td>
<td>Private</td>
<td>HS_grad</td>
<td>9</td>
<td>Never_married</td>
<td>...</td>
<td>Machine_c</td>
<td>Unmarried</td>
<td>White</td>
<td>Male</td>
<td>40</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>58</td>
<td>Private</td>
<td>11th</td>
<td>7</td>
<td>Married</td>
<td>...</td>
<td>Sales</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>50</td>
<td>United_States</td>
<td>poor</td>
</tr>
<tr>
<td>44</td>
<td>Self_emp</td>
<td>Masters</td>
<td>14</td>
<td>Divorced</td>
<td>...</td>
<td>Exec_mani</td>
<td>Unmarried</td>
<td>Female</td>
<td>45</td>
<td>United_States</td>
<td>rich</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>Private</td>
<td>Doctorate</td>
<td>16</td>
<td>Married</td>
<td>...</td>
<td>Prof_speical</td>
<td>Husband</td>
<td>White</td>
<td>Male</td>
<td>60</td>
<td>United_States</td>
<td>rich</td>
</tr>
</tbody>
</table>

molecule: alanine
Contents

Position of the problem
  Background notations
  **Difficulties**
  The learning process
  The villain

Validation
  Performance indicators
  Estimating an indicator
  Testing a hypothesis
  Comparing hypotheses

Validation Campaign
  The point of parameter setting
  Racing
  Expected Global Improvement
Difficulty factors

Quality of examples / of representation

+ Relevant features
  − Not enough data
  − Noise; missing data
  − Structured data: spatio-temporal, relational, textual, videos ..

Distribution of examples

+ Independent, identically distributed examples
  − Other: robotics; data stream; heterogeneous data

Prior knowledge

+ Constraints on sought solution
+ Criteria; loss function
Difficulty factors, 2

Learning criterion

• Convex function: a single optimum

↘ Complexity: $n$, $n \log n$, $n^2$

– Combinatorial optimization

What is your agenda?

▶ Prediction performance
▶ Causality
▶ INTELLIGIBILITY
▶ Simplicity
▶ Stability
▶ Interactivity, visualisation

Scalability
Difficulty factors, 3

Crossing the chasm

- There exists no *killer algorithm*
- Few general recommendations about algorithm selection

Performance criteria

- Consistency

  When number $n$ of examples goes to $\infty$ and the target concept $h^*$ is in $\mathcal{H}$, Algorithm finds $\hat{h}_n$, with

  $$\lim_{n \to \infty} h_n = h^*$$

- Convergence speed

  $$\|h^* - h_n\| = \mathcal{O}(1/n), \mathcal{O}(1/\sqrt{n}), \mathcal{O}(1/\ln n)$$
Contents

Position of the problem
- Background notations
- Difficulties
- The learning process
- The villain

Validation
- Performance indicators
- Estimating an indicator
- Testing a hypothesis
- Comparing hypotheses

Validation Campaign
- The point of parameter setting
- Racing
- Expected Global Improvement
Context

Related approaches

- Data Mining, KDD
  - scalability
- Statistics and data analysis
  - Model selection and fitting; hypothesis testing
- Machine Learning
  - Prior knowledge; representations; distributions
- Optimisation
  - well-posed / ill-posed problems
- Computer Human Interface
  - No ultimate solution: a dialog
- High performance computing
  - Distributed data; privacy
Methodology

Phases

1. Collect data
2. Clean data
3. Select data
4. Data Mining / Machine Learning
   - Description
   - Prediction
   - Aggregate
5. Visualisation
6. Evaluation
7. Collect new data

An interactive process
depending on expectations, data, prior knowledge, current results
Contents

Position of the problem
  Background notations
  Difficulties
  The learning process
  The villain

Validation
  Performance indicators
  Estimating an indicator
  Testing a hypothesis
  Comparing hypotheses

Validation Campaign
  The point of parameter setting
  Racing
  Expected Global Improvement
Supervised Machine Learning

Context

World $\rightarrow$ instance $x_i$ $\rightarrow$ Oracle $\downarrow$ $y_i$

Input
Training set $\mathcal{E} = \{(x_i, y_i), i = 1 \ldots n, x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$

Tasks
- Select hypothesis space $\mathcal{H}$
- Assess hypothesis $h \in \mathcal{H}$ $\text{score}(h)$
- Find best hypothesis $h^*$
What is the point?

Underfitting

Overfitting

The point is not to be perfect on the training set
What is the point?

Underfitting

Overfitting

The point is not to be perfect on the training set

The villain: overfitting

Test error

Training error

Complexity of Hypotheses
Prediction good on future instances

Necessary condition:
Future instances must be similar to training instances
   “identically distributed”

Minimize (cost of) errors
not all mistakes are equal. \[ \ell(y, h(x)) \geq 0 \]
Minimize expectation of error cost

Minimize $E[\ell(y, h(x))] = \int_{X \times Y} \ell(y, h(x))p(x, y)dx \, dy$
Error: theoretical approach

Minimize expectation of error cost

\[ \text{Minimize } E[\ell(y, h(x))] = \int_{X \times Y} \ell(y, h(x)) p(x, y) \, dx \, dy \]

Principle
Si \( h \) “is well-behaved” on \( \mathcal{E} \), and \( h \) is ”sufficiently regular” \( h \) will be well-behaved in expectation.

\[ E[F] \leq \frac{\sum_{i=1}^{n} F(x_i)}{n} + c(F, n) \]
Classification, Problem posed

INPUT

\[ \mathcal{E} = \{(x_i, y_i), x_i \in \mathcal{X}, y_i \in \{0, 1\}, i = 1 \ldots n\} \]

HYPOTHESIS SPACE

\[ \mathcal{H} \quad h : \mathcal{X} \mapsto \{0, 1\} \]

SEARCH SPACE

LOSS FUNCTION

\[ \ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R} \]

OUTPUT

\[ h^* = \arg \max \{\text{score}(h), h \in \mathcal{H}\} \]
Classification, criteria

Generalisation error

\[ Err(h) = E[\ell(y, h(x))] = \int \ell(y, h(x))dP(x, y) \]

Empirical error

\[ Err_e(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(x_i)) \]

Bound

risk minimization

\[ Err(h) < Err_e(h) + F(n, d(\mathcal{H})) \]

\[ d(\mathcal{H}) = \text{VC-dimension of } \mathcal{H} \]
Dimension of Vapnik Cervonenkis

**Principle** Given hypothesis space $\mathcal{H} : \mathcal{X} \mapsto \{0, 1\}$ Given $n$ points $x_1, \ldots, x_n$ in $\mathcal{X}$.

If, $\forall (y_i)_{i=1}^n \in \{0, 1\}^n$, $\exists h \in \mathcal{H} / h(x_i) = y_i,$

$\mathcal{H}$ shatters $\{x_1, \ldots, x_n\}$

Example: $\mathcal{X} = \mathbb{R}^p$

$d(\text{hyperplanes in } \mathbb{R}^p) = p + 1$

WHY: if $\mathcal{H}$ shatters $\mathcal{E}$, $\mathcal{E}$ doesn’t tell anything

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3 pts shattered by a line

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4 points, non shattered

**Definition**

$$d(\mathcal{H}) = \max\{n/\exists(x_1 \ldots, x_n) \text{ shattered by } \mathcal{H}\}$$
Classification: Ingredients of error

Bias
Bias (\(\mathcal{H}\)): error of the best hypothesis \(h^*\) in \(\mathcal{H}\)

Variance
Variance of \(h_n\) depending on \(\mathcal{E}\)

Optimization
negligible in small scale
takes over in large scale

Google
Contents

Position of the problem
  Background notations
  Difficulties
  The learning process
  The villain

Validation
  Performance indicators
  Estimating an indicator
  Testing a hypothesis
  Comparing hypotheses

Validation Campaign
  The point of parameter setting
  Racing
  Expected Global Improvement
Validation: Three questions

Define a good indicator of quality
- Misclassification cost
- Area under the ROC curve

Computing an estimate thereof
- Validation set
- Cross-Validation
- Leave one out
- Bootstrap

Compare estimates: Tests and confidence levels
Which indicator, which estimate: it depends.

Settings
- Large/few data

Data distribution
- Dependent/independent examples
- balanced/imbalanced classes
Contents

Position of the problem
  Background notations
  Difficulties
  The learning process
  The villain

Validation
  Performance indicators
  Estimating an indicator
  Testing a hypothesis
  Comparing hypotheses

Validation Campaign
  The point of parameter setting
  Racing
  Expected Global Improvement
Performance indicators

Binary class
- $h^*$ the truth
- $\hat{h}$ the learned hypothesis

Confusion matrix

<table>
<thead>
<tr>
<th>$\hat{h}$ / $h^*$</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>0</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>$a+c$</td>
<td>$b+d$</td>
</tr>
</tbody>
</table>
Performance indicators, 2

<table>
<thead>
<tr>
<th>$\hat{h} / h^*$</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>0</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>$a + c$</td>
<td>$b + d$</td>
</tr>
</tbody>
</table>

- Misclassification rate $\frac{b + c}{a + b + c + d}$
- Sensitivity, True positive rate (TP) $\frac{a}{a + c}$
- Specificity, False negative rate (FN) $\frac{b}{b + d}$
- Recall $\frac{a}{a + c}$
- Precision $\frac{a}{a + b}$

Note: always compare to random guessing / baseline alg.
Performance indicators, 3

The Area under the ROC curve

- ROC: Receiver Operating Characteristics
- Origin: Signal Processing, Medicine

Principle

\[ h : X \mapsto \mathbb{R} \quad h(x) \text{ measures the risk of patient } x \]

\( h \) leads to order the examples:

+ + + - - + + + + - - - + - - - + - - - - - - - - - - - -
Performance indicators, 3

The Area under the ROC curve

- ROC: Receiver Operating Characteristics
- Origin: Signal Processing, Medicine

Principle

\[ h : X \mapsto \mathbb{R} \quad h(x) \text{ measures the risk of patient } x \]

\( h \) leads to order the examples:

\[ +++-+-+-+++-++++-+-++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-AAAAAA

Given a threshold \( \theta \), \( h \) yields a classifier: Yes iff \( h(x) > \theta \).

\[ +++-+-+-+++-++++-+-++++-++++++-+-++++-++++++-+-++++-++++++-+-++++-AAAAAA

Here, \( TP(\theta) = .8; \) FN \( (\theta) = .1 \)
ROC

Without disease

Fuller of

With
disease

TN

FN

FP

TP
The ROC curve

Ideal classifier: (0 False negative, 1 True positive)
Diagonal (True Positive = False negative) \equiv nothing learned.
ROC Curve, Properties

Properties
ROC depicts the trade-off True Positive / False Negative.

Standard: misclassification cost (Domingos, KDD 99)

Error = # false positive + $c \times$ # false negative

In a multi-objective perspective, ROC = Pareto front.

Best solution: intersection of Pareto front with $\Delta(-c, -1)$
ROC Curve, Properties, foll’d

Used to compare learners
multi-objective-like
insensitive to imbalanced distributions
shows sensitivity to error cost.

Bradley 97
Area Under the ROC Curve

Often used to select a learner
Don’t ever do this! \textit{Hand, 09}

Sometimes used as learning criterion \textit{Mann Whitney Wilcoxon}

\[ AUC = \Pr(h(x) > h(x') | y > y') \]

\textbf{WHY} \textit{Rosset, 04}
\begin{itemize}
  \item More stable $\mathcal{O}(n^2)$ vs $\mathcal{O}(n)$
  \item With a probabilistic interpretation \textit{Clemençon et al. 08}
\end{itemize}

\textbf{HOW} \textit{Joachims 05; Usunier et al. 08, 09}
\begin{itemize}
  \item SVM-Ranking
  \item Stochastic optimization
\end{itemize}
Contents

Position of the problem
  Background notations
  Difficulties
  The learning process
  The villain

Validation
  Performance indicators
  Estimating an indicator
  Testing a hypothesis
  Comparing hypotheses

Validation Campaign
  The point of parameter setting
  Racing
  Expected Global Improvement
Validation, principle

Desired: performance on further instances

Assumption: Dataset is to World, like Training set is to Dataset.
Validation, 2

Unbiased Assessment of Learning Algorithms
T. Scheffer and R. Herbrich, 97
Validation, 2

Unbiased Assessment of Learning Algorithms
T. Scheffer and R. Herbrich, 97
Unbiased Assessment of Learning Algorithms
T. Scheffer and R. Herbrich, 97
Contents

Position of the problem
   Background notations
   Difficulties
   The learning process
   The villain

Validation
   Performance indicators
   Estimating an indicator
   Testing a hypothesis
   Comparing hypotheses

Validation Campaign
   The point of parameter setting
   Racing
   Expected Global Improvement
Confidence intervals

**Definition**
Given a random variable $X$ on $\mathbb{R}$, a $p\%$-confidence interval is $l \subset \mathbb{R}$ such that

$$Pr(X \in l) > p$$

**Binary variable with probability $\epsilon$**
Probability of $r$ events out of $n$ trials:

$$P_n(r) = \frac{n!}{r!(n-r)!} \epsilon^r (1 - \epsilon)^{n-r}$$

- Mean: $n\epsilon$
- Variance: $\sigma^2 = n\epsilon(1 - \epsilon)$

**Gaussian approximation**

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{1}{2} \frac{x-\mu}{\sigma}^2\right)$$
Confidence intervals

Bounds on (true value, empirical value) for $n$ trials, $n > 30$

$$Pr\left( |\hat{x}_n - x^*| > z \cdot 1.96 \cdot \sqrt{\frac{\hat{x}_n(1-\hat{x}_n)}{n}} \right) < \varepsilon$$

<table>
<thead>
<tr>
<th>$\varepsilon$</th>
<th>50</th>
<th>32</th>
<th>20</th>
<th>10</th>
<th>5</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z$</td>
<td>.67</td>
<td>1</td>
<td>1.28</td>
<td>1.64</td>
<td>1.96</td>
<td>2.33</td>
<td>2.58</td>
</tr>
</tbody>
</table>
Empirical estimates

When data abound

(MNIST)

Training  Test  Validation

Cross validation

Fold

Run

N–fold Cross Validation

Error = Average (error on learned from )
Empirical estimates, foll’d

Cross validation $\rightarrow$ Leave one out

![Diagram of cross validation and leave one out]

Same as N-fold CV, with N = number of examples.

Properties

Low bias; high variance; underestimate error if data not independent
Empirical estimates, foll’d

Dataset

Training set

Test set.

rest of examples

with replacement

uniform sampling

Average indicator over all (Training set, Test set) samplings.
Contents

Position of the problem
  Background notations
  Difficulties
  The learning process
  The villain

Validation
  Performance indicators
  Estimating an indicator
  Testing a hypothesis
  Comparing hypotheses

Validation Campaign
  The point of parameter setting
  Racing
  Expected Global Improvement
Is \( \hat{h} \) better than random?

The McNemar test

<table>
<thead>
<tr>
<th>( \hat{h} / h^* )</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>0</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>( a+c )</td>
<td>b+d</td>
<td>a + b + c + d</td>
</tr>
</tbody>
</table>

Property \( \frac{|b-c|-1}{b+c} \) follows a \( \chi^2 \) law with degree of freedom 1.
Types of test error

Type I error
The hypothesis is not significant, and the test thinks it’s significant

Type II error
The hypothesis is valid, and the test discards it.
Comparing algorithms A and B

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>A-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>run 1</td>
<td>30</td>
<td>28</td>
<td>2</td>
</tr>
<tr>
<td>run 2</td>
<td>17</td>
<td>25</td>
<td>-8</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>25</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>28</td>
<td>-11</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>26</td>
<td>4</td>
</tr>
</tbody>
</table>

Assumption
A and B have normal distribution

Simplest case
two samples with same size, (quasi)
same variance.

Define

\[ t = \frac{\bar{A} - \bar{B}}{S_{A,B} \cdot \sqrt{\frac{2}{n}}} \]

with \( S_{A,B} = \sqrt{\frac{1}{2}(S_A^2 + S_B^2)} \) and \( S_A^2 = \frac{1}{n} \sum (A_i - \bar{A})^2 \)
Comparing algorithms A and B

\( t \) follows a Student law with \((2n-2)\)-dof

- Compute \( t \)
- See confidence of \( t \)
Comparing algorithms A and B

Recommended: Use paired t-test

- Apply A and B with same (training, test) sets
- Variance is lower:

\[ Var(A - B) = Var(A) + Var(B) - 2\text{coVar}(A, B) \]

- Thus easier to make significant differences

What if variances are different?
See Welch’ test:

\[ \frac{\bar{A} - \bar{B}}{\sqrt{\frac{S^2_A}{N_A} + \frac{S^2_B}{N_B}}} \]
Summary: single dataset (if we had enough data...)

The 5 x 2CV

- 5 times
- split the data into 2 halves
- gives 10 estimates of error indicator
  + More independent
  - Each training set is 1/2 data.

With a single dataset

- 5x2 CV
- paired t-test
- McNemar test on a validation set

Dietterich 98
Multiple datasets

If A and B results don’t follow a normal distribution

\[ Z_i = A_i - B_i \]

Wilcoxon signed rank test

1. Rank the \(|Z_i|\)
2. \(W_+ = \text{sum of ranks when } Z_i > 0\)
3. \(W_- = \text{sum of ranks when } Z_i < 0\)
4. \(W_{\text{min}} = \text{min}(W_+, W_-)\)

\[ z = \frac{1/4n(n+1) - W_{\text{min}} - 1/2}{\sqrt{1/24n(n+1)(2n+1)}} \]

5. \(z \sim \mathcal{N}(0, 1) \quad n > 20\)

| A  | B  | |Z| | rank | sign |
|----|----|---|----|-----|-----|
| 19 | 23 | 4 | 6th | −   |
| 22 | 21 | 1 | 1st | +   |
| 21 | 19 | 2 | 2nd | +   |
| 25 | 28 | 3 | 4th | −   |
| 24 | 22 | 2 | 2nd | +   |
| 23 | 20 | 3 | 4th | +   |
Multiple hypothesis testing

Beware

- If you test many hypotheses on the same dataset
- one of them will appear confidently true...
  increase in type I error

Corrections Over $n$ tests, the global significance level $\alpha_{\text{global}}$ is related to the elementary significance level $\alpha_{\text{unit}}$:

$$\alpha_{\text{global}} = 1 - (1 - \alpha_{\text{unit}})^n$$

- Bonferroni correction pessimistic

$$\alpha_{\text{unit}} = \frac{\alpha_{\text{global}}}{n}$$

- Sidak correction

$$\alpha_{\text{unit}} = 1 - (1 - \alpha_{\text{global}})^{\frac{1}{n}}$$
Contents

Position of the problem
  Background notations
  Difficulties
  The learning process
  The villain

Validation
  Performance indicators
  Estimating an indicator
  Testing a hypothesis
  Comparing hypotheses

Validation Campaign
  The point of parameter setting
  Racing
  Expected Global Improvement
How to set up my system?

Parameter tuning
- Setting the parameters for feature extraction
- Select the best learning algorithm
- Setting the learning parameters (e.g. type of kernel, the parameters in SVMs)
- Setting the validation parameters

Goal: find the best setting
- Algorithm selection in Operational Research
- Parameter tuning in Stochastic Optimization
- Meta-Learning in Machine Learning

a pervasive concern
Main approaches

1. Design of experiments (Latin square)
2. Anova (Analysis of variance)-like methods:
   - Racing
   - Sequential parameter optimization
Parameter Tuning: A Meta-Optimization problem

Optimization: the Black-Box Scenario

- Need to perform several runs to compute performance
- Need to specify the number of runs
- Overall cost is the total number of evaluations
- And don’t forget to tune the parameters of the meta-optimizer!
Parameter Tuning: A Meta-Optimization problem

Learning & valid. parameters

Learner | Dataset
--------|--------

Validation

performance

Optimization: the Black-Box Scenario

- Need to perform several runs to compute performance using Cross-Validation
- Need to specify the number of runs and tune it optimally
- Overall cost is the total number of evaluations
- And don’t forget to tune the parameters of the meta-optimizer!
Parameter Tuning: A Meta-Optimization problem

Optimization: the Black-Box Scenario

- Need to perform several runs to compute performance
- Need to specify the number of runs
- Overall cost is the total number of evaluations
- And don’t forget to tune the parameters of the meta-optimizer!
Ingredients

Design Of Experiments (DOE)
- A long-known method from statistics
- Choose a finite number of parameter sets
- Compute their performance
- Return the statistically significantly best sets

Analysis of Variance (ANOVA)
- Assumes normally distributed data
- Tests if means are significantly different for a given confidence level; generalizes T-Test
- Perform pairwise tests if ANOVA reports some difference
  T-Test, rank-based tests, ...
DOE: Issues

Choice of sample parameter sets

- **Full Factorial Design**
  - Discretize all parameters if continuous
  - Choose all possible combinations

- **Latin Hypercube Sampling**: to generate $k$ sets,
  - Discretize all parameters in $k$ values
  - Repeat $k$ times:
    - for each parameter, (uniformly) choose one value out of $k$
  - For each parameter, each value is taken once
    - fine if no correlation

Cost

- For each parameter set, the full cost of learning validation
- Combinatorial explosion with number of parameters and precision
Racing algorithms

Birattari & al. 02, Yuan & Gallagher 04

Rationale

▷ All parameter settings are run the same number of times whereas very bad settings could be detected earlier

Implementation

▷ Repeat
  ▷ Perform only a few runs per parameter set
  ▷ Statistically check all sets against the best one at given confidence level
  ▷ Discard the bad ones
▷ Until only survivor, or maximum number of runs per setting reached
Racing algorithms

How?

Example: Initialization

- $R = 0$
- while $R < R_{max}$ and more than 1 set
  - Compute empirical value of performance for all sets doing $r$ additional runs
    - average, median, . . .
  - Compute $X\%$ confidence intervals
    - Hoeffding bounds, Friedman tests, . . .
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
- $R+ = r$
Racing algorithms

How?

Example: Initialization

- \( R = 0 \)
- while \( R < R_{\text{max}} \) and more than 1 set
  - Compute empirical value of performance for all sets doing \( r \) additional runs
    - average, median, \ldots
  - Compute X\% confidence intervals
    - Hoeffding bounds, Friedman tests, \ldots
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
- \( R+ = r \)
Racing algorithms

How?

Example: Initialization

- \( R = 0 \)
- while \( R < R_{\text{max}} \) and more than 1 set
  - Compute empirical value of performance for all sets doing \( r \) additional runs average, median, . . .
  - Compute \( X\% \) confidence intervals Hoeffding bounds, Friedman tests, . . .
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
- \( R+ = r \)
Racing algorithms

Example: Iteration 1

How?

- $R = 0$
- while $R < R_{\text{max}}$ and more than 1 set
  - Compute empirical value of performance for all sets doing $r$ additional runs
    - average, median, . . .
  - Compute $X\%$ confidence intervals
    - Hoeffding bounds, Friedman tests, . . .
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
- $R+ = r$
Racing algorithms

How?

Example: Iteration 1

- \( R = 0 \)
- while \( R < R_{max} \) and more than 1 set
  - Compute empirical value of performance for all sets doing \( r \) additional runs average, median, . . .
  - Compute \( X\% \) confidence intervals Hoeffding bounds, Friedman tests, . . .
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
- \( R+ = r \)
Racing algorithms

How?

Example: Iteration N

- $R = 0$
- while $R < R_{\text{max}}$ and more than 1 set
  - Compute empirical value of performance for all sets doing $r$ additional runs average, median, ...
  - Compute $X\%$ confidence intervals
    - Hoeffding bounds, Friedman tests, ...
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
- $R+ = r$
Racing algorithms

How?

Example: Iteration N

- $R = 0$
- while $R < R_{\text{max}}$ and more than 1 set
  - Compute empirical value of performance for all sets doing $r$ additional runs average, median, . . .
  - Compute $X\%$ confidence intervals Hoeffding bounds, Friedman tests, . . .
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
- $R+ = r$
Racing algorithms

**How?**

- $R = 0$
- while $R < R_{\text{max}}$ and more than 1 set
  - Compute empirical value of performance for all sets doing $r$ additional runs
  - average, median, . . .
  - Compute X% confidence intervals Hoeffding bounds, Friedman tests, . . .
  - Remove sets whose best possible value is worse than worse possible value of the best empirical set.
- $R^+ = r$

**Example:** Best parametere sets
Racing algorithms: Discussion

Results

- Published results claim saving between 50 and 90% of the runs

Useful for

- Multiple algorithms on single problem for efficiency
- Single algorithm on multiple problems to assess problem difficulties
- Multiple algorithms on multiple problems for robustness

Issues

- Nevertheless costly
- Can only find the best one in initial sample
Rationale

- Start with some very coarse sampling DOE
- Evaluate performance using few runs per set
- Build a model of the performance landscape using Gaussian Processes aka Kriging
- Select best points based on Expected Improvement according to current model Monte-Carlo sampling
- Compute actual performance of best estimates using same number of runs as current best
- Increase # runs of best if unchanged
Gaussian Processes in one slide

An optimization algorithm for expensive functions

D.R. Jones, Schonlau, & Welch, 98
Gaussian Processes in one slide
An optimization algorithm for expensive functions

D.R. Jones, Schonlau, & Welch, 98
Gaussian Processes in one slide
An optimization algorithm for expensive functions

D.R. Jones, Schonlau, & Welch, 98
Pros

- Similar ideas as racing,
- but allows to refine initial sampling a true optimization algorithm
- Compatible with a fixed budget scenario racing is not
- Authors also report gains up to 90%

Cons

- Works best with ... some tuning
Take home messages

What is the performance criterion

- Cost function
- Account for class imbalance
- Account for data correlations

Assessing a result

- Compute confidence intervals
- Consider baselines
- Use a validation set

If the result looks too good, beware