Text Classification, K-NN, Evaluation

Wm&R a.a. 2009/10

R. Basili

Dipartimento di Informatica Sistemi e produzione
Università di Roma “Tor Vergata”
Email: basili@info.uniroma2.it
Text Categorization and Optimization
- TC Introduction
- TC: Performance Evaluation
- The designing steps of a TC system
- The Rocchio classifier
- Lazy Learning: K-NN
- The Parameterized Rocchio Classifier (PRC)
- Comparative Evaluation: Rocchio, PRC and SVM
Introduction to Text Categorization

Berlusconi acquires Inzaghi before elections

Politic \( C_1 \)
Economic \( C_2 \)
Sport \( C_n \)
Text Classification Problem

- Given:
  - a set of target categories: \( C = \{ C^1, \ldots, C^n \} \)
  - the set \( T \) of documents,

  define
  \[
  f : T \rightarrow 2^C
  \]

- VSM (Salton89’)
  - Features are dimensions of a Vector Space.
  - Documents and Categories are vectors of feature weights.
  - \( d \) is assigned to \( C^i \) if
    \[
    \vec{d} \times \vec{C}^i > th
    \]
The Vector Space Model

- **$d_1$: Politic**
  - Bush declares war.
  - Berlusconi gives support

- **$d_2$: Sport**
  - Wonderful Totti in the yesterday match against Berlusconi’s Milan

- **$d_3$: Economic**
  - Berlusconi acquires Inzaghi before elections

**Categories**
- $C_1$: Politics Category
- $C_2$: Sport Category
Automated Text Categorization

- A corpus of pre-categorized documents
- Split document in two parts:
  - Training-set
  - Test-set
- Apply a supervised machine learning model to the training-set
  - Positive examples
  - Negative examples
- Measure the performances on the test-set
  - e.g., Precision and Recall
Performance Measurements

- Given a set of document $T$
- Precision = $\frac{\text{# Correct Categories Retrieved}}{\text{# Categories Retrieved}}$
- Recall = $\frac{\text{# Correct Categories Retrieved}}{\text{# Correct Categories (by the system)}}$

![Diagram showing the intersection of Categories Retrieved (by the system) and Correct Categories, with the overlap representing the Correct Categories Retrieved (by the system).]
Precision and Recall of $C_i$

- $a$, numb. of correct labeling/classification
- $b$, numb. of mistakes, wrong labeling
- $c$, numb. of not retrieved labels

The *Precision* and *Recall* are defined by the above counts:

\[
\text{Precision}_i = \frac{a_i}{a_i + b_i}
\]

\[
\text{Recall}_i = \frac{a_i}{a_i + c_i}
\]
Breakeven Point
- Find thresholds for which \( \text{Recall} = \text{Precision} \)
- Interpolation

\textbf{f-measure}
- Harmonic mean between precision and recall

Global performance on more than two categories
- Micro-average
  - The counts refer to classifiers
- Macro-average (average measures over all categories)
F-measure e MicroAverages

\[ F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]

\[ \mu\text{Precision} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} a_i + b_i} \]

\[ \mu\text{Recall} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} a_i + c_i} \]

\[ \mu\text{BEP} = \frac{\mu\text{Precision} + \mu\text{Recall}}{2} \]

\[ \mu f_1 = \frac{2 \times \mu\text{Precision} \times \mu\text{Recall}}{\mu\text{Precision} + \mu\text{Recall}} \]
N-fold cross validation

- In order to obtain a more stable measure for F1, it is natural to repeat the evaluation over multiple splitting.

- For each split, the entire oracle (i.e. the gold standard of labeled examples) is divided into $n$ parts:
  - One part is used for testing.
  - The remaining $n-1$ parts are used for training the classifier.

- The final result is the mean (expected) value of F1 plus/minus the standard deviation obtained over the $n$ repeated measures.
Text Categorization phases

- Corpus pre-processing (e.g. tokenization, stemming)
- Feature Selection (optionally)
  - Document Frequency, Information Gain, $\chi^2$, mutual information,...
- Feature weighting
  - for documents and profiles
- Similarity measure
  - between document and profile (e.g. scalar product)
- Statistical Inference
  - threshold application
- Performance Evaluation
  - Accuracy, Precision/Recall, BEP, f-measure,...
Feature Selection

- Un certo numero di feature potrebbero non essere rilevanti.
- Per esempio, in TC le “function words” come “the”, “on”, “those”...
- L’algoritmo di learning ha due benefici:
  - Migliora l’efficienza
  - Migliora l’accuratezza
- Ordinare le feature per rilevanza e prendere le $m$ più rilevanti.
Statistical Quantity to sort feature

- Based on corpus counts of the pair <feature, category>

- $A$ is the number of documents in which both $f$ and $c$ occur, i.e. $(f, c)$;
- $B$ is the number of documents in which only $f$ occurs, i.e. $(f, \bar{c})$;
- $C$ is the number of documents in which only $c$ occurs, i.e. $(\bar{f}, c)$;
- $D$ is the number of documents in which neither $f$ nor $c$ occur, i.e. $(\bar{f}, \bar{c})$;
- $N$ is the total number of documents, i.e. $A + B + C + D$. 
Statistical Selectors

- Chi-square, pointwise MI $e$ MI

\[ \chi^2(f, c) = \frac{N \times (AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)} \]

\[ PMI(f, c) = \log \frac{P(f, c)}{P(f) \times P(c)} \]

\[ MI(f) = - \sum_{c \in C} P(c) \log(P(c)) + P(f) \sum_{c \in C} P(c|f) \log(P(c|f)) + P(\bar{f}) \sum_{c \in C} P(c|\bar{f}) \log(P(c|\bar{f})) \]
- \( P(f, c) \) is the probability that \( f \) and \( c \) co-occurs and can be estimated by \( A/N \);

- \( P(f) \) is the probability of \( f \), estimated by \( (A + B)/N \);

- \( P(c) \) is the probability of \( f \), estimated by \( (A + C)/N \);

- \( P(c|f) \) is the probability of \( c \) by considering only the documents that contain \( f \). It can be estimated by \( \frac{P(f,c)}{P(f)} \).

- \( P(\overline{f}) \) is the probability that \( f \) does not occur, estimated by \( (C + D)/N \);
Probability Estimation (con’t)

- $P(c|\bar{f})$ is the probability of $c$ by considering only the documents that do not contain $f$. It can be estimated by $\frac{P(\bar{f},c)}{P(f)}$. In turn, $P(\bar{f},c)$ is estimated by $C/N$.

- $C$ is the collection of categories, i.e. $\{c_1, c_2, \ldots, c_n\}$. Note that $PMI$ and $\chi^2$ are defined on only two categories, i.e. $c$ and not $c$ whereas $MI$ can be evaluated on $n > 2$ categories\(^7\).

For example, we can apply the above formulas to evaluate the $PMI$ as follows:

$$PMI(f,c) = \frac{A \times N}{(A + C)(A + B)}$$
Global Selectors

\[ PMI_{max}(f) = \max_{c \in \mathcal{C}} \ MI(f, c) \]

\[ PMI_{avg}(f) = \sum_{c \in \mathcal{C}} P(c) \times MI(f, c) \]

\[ \chi^2_{max}(f) = \max_{c \in \mathcal{C}} \ \chi^2(f, c) \]

\[ \chi^2_{avg}(f) = \sum_{c \in \mathcal{C}} P(c) \times \chi^2(f, c) \]
Document weighting: an example

- $N$, the overall number of documents,
- $N_f$, the number of documents that contain the feature $f$
- $o_f^d$, the occurrences of the features $f$ in the document $d$

The weight $f$ in a document is:

$$w_f^d = (\log \frac{N}{N_f}) o_f^d = IDF(f) \cdot o_f^d$$

The weight can be normalized:

$$w'_f^d = \frac{w_f^d}{\sqrt{\sum_{i \in d} (w_i^d)^2}}$$
Similarity estimation

- Given the document representation
  \[ \vec{d} = \langle w_{f_1}^d, \ldots, w_{f_n}^d \rangle \]

- Given the category representation
  \[ \vec{C}_i = \langle W_{f_1}^i, \ldots, W_{f_n}^i \rangle \]

- It can be defined the following similarity function (cosine measure)
  \[ s_{di} = \cos(\angle \vec{d} , \vec{C}_i) = \frac{\vec{d} \times \vec{C}_i}{\| \vec{d} \| \times \| \vec{C}_i \|} = \frac{\sum_{f} w_{f}^d \times W_{f}^i}{\| \vec{d} \| \times \| \vec{C}_i \|} \]
The Rocchio Classifier

- $w_f^d$, the weight of $f$ in $d$
  - Several weighting schemes (e.g. TF * IDF, Salton 91’)

- $W_f^i$, the profile weights of $f$ in $C_i$:
  
  $$W_f^i = \max \left\{ 0, \frac{\beta}{|T_i|} \sum_{d \in T_i} w_f^d - \frac{\gamma}{|T_i|} \sum_{d \in T_i} w_f^d \right\}$$

- $T_i$, the training documents in $C_i$

- $d$ is assigned to $C_i$ if $\vec{d} \times \vec{C}^i > th$
Bidimensional view of Rocchio categorization
Limitations of the Rocchio TC

- Prototype-based models have problems with polymorphic (i.e. disjunctive) categories.
Feature Selection interpretation of Rocchio parameters

- Literature work uses a bunch of values for $\beta$ and $\gamma$
- Interpretation of positive ($\beta$) vs. negative ($\gamma$) information
  - $\Rightarrow$ value of $\beta > \gamma$ (e.g. 16, 4)

- Our interpretation [IJAIT 2002, ECIR 2003]:
  - Remove one parameters
    \[
    \tilde{C}_f^i = \max \left\{ \frac{1}{|T_i|} \sum_{d \in T_i} \tilde{d}_f - \frac{\rho}{|T_i|} \sum_{d \in T_i} \tilde{d}_f \right\}
    \]
  - 0 weighted features do not affect similarity estimation
  - increasing $\rho$ causes many feature to be set to 0 $\Rightarrow$ they are removed
Feature Selection interpretation of Rocchio parameters (cont’d)

- By increasing $\rho$:
  - Features that have a high negative weights get firstly a zero value
  - High negative weight means very frequent in the other categories
  - $\Rightarrow$ zero weight for irrelevant features

- If $\rho$ is a feature selector, set it according to standard feature selection strategies [Yang, 97]

- Moreover, we can find a maximal value $\rho_{\text{max}}$ (associated with all feature removed)

- This interpretation enabled $\gamma \gg \beta$
Nearest-Neighbor Learning

- Memory-based learning: learning is just storing the representations of the training examples in $D$.

- Testing instance $x$:
  - Compute similarity between $x$ and all examples in $D$.
  - Assign $x$ the category of the most similar example in $D$.

- Does not explicitly compute a generalization or category prototypes.

- Also called:
  - Case-based
  - Memory-based
  - Lazy learning
Using only the closest example to determine categorization is subject to errors due to:

- A single atypical example.
- Noise (i.e. error) in the category label of a single training example.

More robust alternative is to find the $k$ most-similar examples and return the majority category of these $k$ examples.

Value of $k$ is typically odd, 3 and 5 are most common.
5 Nearest Neighbor Illustration (Euclidian Distance)
K Nearest Neighbor for Text

Training:
For each training example \( <x, c(x)> \in D \)
Compute the corresponding TF-IDF vector, \( d_x \), for document \( x \)

Test instance \( y \):
Compute TF-IDF vector \( d \) for document \( y \)
For each \( <x, c(x)> \in D \)
Let \( s_x = \text{cosSim}(d, d_x) \)
Sort examples, \( x \), in \( D \) by decreasing value of \( s_x \)
Let \( N \) be the closest (i.e. first) \( k \) examples in \( D \). \((get \ k \ most \ similar \ neighbors)\)
Return the majority class of examples in \( N \)
Illustration of 3 Nearest Neighbor for Text
Other text classifiers

- **RIPPER** [Cohen and Singer, 1999] uses an extended notion of a profile. It learns the contexts that are positively correlated with the target classes, i.e. words co-occurrence.

- **EXPERT** uses as context nearby words (sequence of words).

- **CLASSI** is a system that uses a neural network-based approach to text categorization [Ng et al., 1997]. The basic units of the network are only perceptrons.

- **Dtree** [Quinlan, 1986] is a system based on a well-known machine learning model.

- **CHARADE** [I. Moulinier and Ganascia, 1996] and **SWAP1** [Apt´e et al., 1994] use machine learning algorithms to inductively extract Disjunctive Normal Form rules from training documents.
Experiments

- Reuters Collection 21578 Apté split (Apté94)
  - 90 classes (12,902 docs)
  - A fixed splitting between training and test set
  - 9603 vs 3299 documents

- Tokens
  - about 30,000 different

- Other different versions have been used but …
  - most of TC results relate to the 21578 Apté
    - [Joachims 1998], [Lam and Ho 1998], [Dumais et al. 1998],
      [Li Yamanishi 1999], [Weiss et al. 1999],
      [Cohen and Singer 1999]…
CRA SOLD FORREST GOLD FOR 76 MLN DLRS - WHIM CREEK

SYDNEY, April 8 - <Whim Creek Consolidated NL> said the consortium it is leading will pay 76.55 mln dlr$ for the acquisition of CRA Ltd's <CRAA.S> <Forrest Gold Pty Ltd> unit, reported yesterday.

CRA and Whim Creek did not disclose the price yesterday.

Whim Creek will hold 44 pct of the consortium, while <Austwhim Resources NL> will hold 27 pct and <Croesus Mining NL> 29 pct, it said in a statement.

As reported, Forrest Gold owns two mines in Western Australia producing a combined 37,000 ounces of gold a year. It also owns an undeveloped gold project.
WASHINGTON, March 20 - The Federal Trade Commission said its staff has urged the governor of Georgia to veto a bill that would prohibit petroleum refiners from owning and operating retail gasoline stations.

The proposed legislation is aimed at preventing large oil refiners and marketers from using predatory or monopolistic practices against franchised dealers.

But the FTC said fears of refiner-owned stations as part of a scheme of predatory or monopolistic practices are unfounded. It called the bill anticompetitive and warned that it would force higher gasoline prices for Georgia motorists.
Feature Selection interpretation of Rocchio parameters

- Literature work uses a bunch of values for \( \beta \) and \( \gamma \)
- Interpretation of positive (\( \beta \)) vs. negative (\( \gamma \)) information
  \[ \Rightarrow \text{value of } \beta > \gamma \text{ (e.g. 16, 4)} \]
- Our interpretation [IJAIT 2002, ECIR 2003]:
  - Remove one parameters
  \[
  \tilde{C}^i_f = \max \left\{ 0, \frac{1}{|T_i|} \sum_{d \in T_i} \tilde{d}_f - \frac{\rho}{|T_i|} \sum_{d \in \overline{T_i}} \tilde{d}_f \right\}
  \]
  - 0 weighted features do not affect similarity estimation
  - increasing \( \rho \) causes many feature to be set to 0 \( \Rightarrow \) they are removed
The Impact of $\rho$ parameter on Acquisition category
The impact of $\rho$ parameter on Trade category

BEP

$\rho$
Mostly populated categories

![Graph with categories and values]
Medium sized categories

![Graph showing BEP, Trade, Interest, and Money-Supply categories.](image-url)
Low size categories
Parameter Estimation Procedure

- Validation-set of about 30% of the training corpus
- for all $\rho \in [0,30]$
  - TRAIN the system on the remaining material
  - Measure the BEP on the validation-set
- Pick-up the $\rho$ associated to the highest $BEP$
- re-TRAIN the system on the entire training-set
- TEST the system based on the obtained parameterized model
- For more reliable results:
  - 20 validation-sets and made the $\rho$ average
- The Parameterized Rocchio Classifier will refer to as PRC
Comparative Analysis

- Rocchio literature parameterization
  - \( \rho = 1 \ (\gamma = \beta = 1) \) and \( \rho = \frac{1}{4} \ (\gamma = 4, \ \beta = 16) \)
- Reuters fixed test-set
  - Other literature results
- SVM
  - To better collocate our results
- Cross Validation (20 samples)
  - More reliable results
- Cross corpora/language validation
  - Reuters, Ohsumed (English) and ANSA (Italian)
## Results on Reuters fixed split

<table>
<thead>
<tr>
<th>Feature Set (~30.000)</th>
<th>PRC</th>
<th>Std Rocchio (γ = ¼ β or γ = β)</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens</td>
<td>82.83 %</td>
<td>72.71% - 78.79 %</td>
<td>85.34 %</td>
</tr>
<tr>
<td>Literature (stems)</td>
<td>-</td>
<td>75 % - 79.9%</td>
<td>84.2 %</td>
</tr>
</tbody>
</table>

- Rocchio literature results (Yang 99’, Choen 98’, Joachims98’)
- SVM literature results (Joachims 98’)

---

- PRC: Precision-Recall Curve
- Std Rocchio: Standard Rocchio
- SVM: Support Vector Machine
Breakeven points of widely known classifiers on Reuters

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>PRC</th>
<th>KNN</th>
<th>RIPPER</th>
<th>CLASSI*</th>
<th>Dtree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>85.34%</td>
<td>82.83%</td>
<td>82.3%</td>
<td>82%</td>
<td>80.2%</td>
<td>79.4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SWAP1*</th>
<th>CHARADE*</th>
<th>EXPERT</th>
<th>Rocchio</th>
<th>Naive Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80.5%</td>
<td>78.3%</td>
<td>82.7%</td>
<td>72%-79.5%</td>
<td>75 % - 79.9%</td>
</tr>
</tbody>
</table>

* Evaluation on different Reuters versions
Cross-Validation

1. Generate $n$ random splits of the corpus. For each split $j$, 70% of data can be used for training ($LS^j$) and 30% for testing ($TS^j$).

2. For each split $j$
   (a) Generate $m$ validation sets, $ES^j_k$ of about 10/30% of $LS^j$.
   (b) Learn the classifiers on $LS^j - ES^j_k$ and for each $ES^j_k$ evaluate: (i) the threshold associated to the BEP and (ii) the optimal parameter $\rho$.
   (c) Learn the classifiers Rocchio, SVMs and PRC on $LS^j$: in case of PRC use the estimated $\bar{\rho}$.
   (d) Evaluate $f_1$ on $TS_j$ (use the estimated thresholds for Rocchio and PRC) for each category and account data for the final processing of the global $\mu f_1$.

3. For each classifier evaluate the mean and the Standard Deviation for $f_1$ and $\mu f_1$ over the $TS_j$ sets.
# Cross-Validation on Reuters (20 samples)

<table>
<thead>
<tr>
<th></th>
<th>Rocchio</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RTS ρ=.25</td>
<td>TS ρ=1</td>
<td>RTS ρ=.25</td>
<td>TS ρ=1</td>
<td></td>
</tr>
<tr>
<td>earn</td>
<td>95.69</td>
<td>95.61</td>
<td>92.57 ±0.51</td>
<td>93.71 ±0.42</td>
<td>95.31</td>
</tr>
<tr>
<td>acq</td>
<td>59.85</td>
<td>82.71</td>
<td>60.02 ±1.22</td>
<td>77.69 ±1.15</td>
<td>85.95</td>
</tr>
<tr>
<td>money-fx</td>
<td>53.74</td>
<td>57.76</td>
<td>67.38 ±2.84</td>
<td>71.60 ±2.78</td>
<td>62.31</td>
</tr>
<tr>
<td>grain</td>
<td>73.64</td>
<td>80.69</td>
<td>70.76 ±2.05</td>
<td>77.54 ±1.61</td>
<td>89.12</td>
</tr>
<tr>
<td>crude</td>
<td>73.58</td>
<td>80.45</td>
<td>75.91 ±2.54</td>
<td>81.56 ±1.97</td>
<td>81.54</td>
</tr>
<tr>
<td>trade</td>
<td>53.00</td>
<td>69.26</td>
<td>61.41 ±3.21</td>
<td>71.76 ±2.73</td>
<td>80.33</td>
</tr>
<tr>
<td>interest</td>
<td>51.02</td>
<td>58.25</td>
<td>59.12 ±3.44</td>
<td>64.05 ±3.81</td>
<td>70.22</td>
</tr>
<tr>
<td>ship</td>
<td>69.86</td>
<td>84.04</td>
<td>65.93 ±4.69</td>
<td>75.33 ±4.41</td>
<td>86.77</td>
</tr>
<tr>
<td>wheat</td>
<td>70.23</td>
<td>74.48</td>
<td>76.13 ±3.53</td>
<td>78.93 ±3.00</td>
<td>84.29</td>
</tr>
<tr>
<td>corn</td>
<td>64.81</td>
<td>66.12</td>
<td>66.04 ±4.80</td>
<td>68.21 ±4.82</td>
<td>89.91</td>
</tr>
<tr>
<td>MicroAvg.</td>
<td>72.61</td>
<td>78.79</td>
<td>73.87 ±0.51</td>
<td>78.92 ±0.47</td>
<td>82.83</td>
</tr>
</tbody>
</table>
Ohsumed and ANSA corpora

- **Ohsumed:**
  - Including 50,216 medical abstracts.
  - The first 20,000 documents year 91,
  - 23 *MeSH* diseases categories [Joachims, 1998]

- **ANSA:**
  - 16,000 news items in Italian from the ANSA news agency.
  - 8 target categories,
  - 2,000 documents each,
  - e.g. Politics, Sport or Economics.

- **Testing 30 %**
Replacement of an aortic valve cusp after neonatal endocarditis. Septic arthritis developed in a neonate after an infection of her hand. Despite medical and surgical treatment endocarditis of her aortic valve developed and the resultant regurgitation required emergency surgery. At operation a new valve cusp was fashioned from preserved calf pericardium. Nine years later she was well and had full exercise tolerance with minimal aortic regurgitation.
## Cross validation on Ohsumed/ANSA (20 samples)

<table>
<thead>
<tr>
<th></th>
<th>Rocchio</th>
<th>PRC</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ohsumed</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEP</td>
<td></td>
<td>f1</td>
<td>f1</td>
</tr>
<tr>
<td>MicroAvg.</td>
<td>$\rho=.25$</td>
<td>$\rho=1$</td>
<td></td>
</tr>
<tr>
<td>(23 cat.)</td>
<td>54.4 ± .5</td>
<td>61.8 ± .5</td>
<td>65.8±.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ANSA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEP</td>
<td></td>
<td>f1</td>
<td></td>
</tr>
<tr>
<td>MicroAvg.</td>
<td>$\rho=.25$</td>
<td>$\rho=1$</td>
<td></td>
</tr>
<tr>
<td>(8 cat.)</td>
<td>61.76 ± .5</td>
<td>67.23 ± .5</td>
<td>71.00 ± .4</td>
</tr>
</tbody>
</table>
Computational Complexity

**PRC**
- Easy to implement
- Low training complexity: $O(n \times m \log n \times m)$
  - $(n = $ number of doc and $m = $ max num of features in a document)
- Low classification complexity:
  $\min\{O(M), O(m \times \log(M))\}$ ($M$ is the max numb. of features in a profile)
- Good performances: the second top accurate classifier on Reuters

**SVM**
- More complex implementation
- Higher Learning time $> O(n^2)$ (to solve the quadratic optimization problem)
- Low complexity of classification phase (for linear SVM) = $\min\{O(M), O(m \times \log(M))\}$
Summary

- The performance evaluation of text classifiers is run against a portion of the gold standard called test set.
- Performance indexes are usually produced on a per class basis \((p_i, r_i, F_i)\) and then they can be globally computed through micro-averaging across classes.
- In this lessons we discussed two geometrical approaches to automatic text classification:
  - Rocchio classifier
    - Non-parametric classifier based on empirical criteria
    - Parametrized version PRC optimize the role of negative examples across classes
    - Good performances: the second top accurate classifier on Reuters
  - Lazy learning: K-NN
    - No generalization is attempted
    - Low complexity if the search for the best \(k\) examples is optimized