Feature Selection

J. Savoy
University of Neuchatel

T. Hastie, R. Tibshirani, J. Friedman: *The Elements of Statistical Learning*. Springer, New York, 2009.

David W. Hosmer, Stanley Lemeshow: *Applied Logistic Regression*. John Wiley & Cons, New York, 2000

Overview

- **The Problem**
- Text Classification
- Binary utility functions
- Ranking the features

Problem

- In classification, we need to have features to predict the corresponding target class
- In regression, features are used to derive the correct target value
- Features are selected to improve the quality of the fit. Each selected feature must be useful to predict the target value (of one or more classes)
- So it it interesting to remove
  - Noisy features (cannot predict the target value)
  - Irrelevant features (can be removed without modifying the performance)
  - Redundant features

Problem

- Working with a reduced number of features may
  - Reduce the storage needed
  - Reduce the learning cost (efficiency)
  - Reduce the risk of overfitting
  - May improve the effectiveness (quality of the prediction)
- Many good reasons to reduce the number of features

Problem

- Why working with less features?
- We want to classify a new instance. We have *p* features and one instance = one point in a *p* dimensional space
- We want a dense representation in the feature space (e.g., every 0.1 distance)
  - If $p = 1$, we need 10 examples
  - If $p = 2$, we need 100 examples
  - If $p = 3$, we need 1,000 examples
  - In general, we need 10^p examples
- Having more features increases the number of training examples

Curse of Dimensionality

- Having more features increases the number of training examples
- But *not* in a *linear* way
  From p to $2p$, we don't double the number of needed training examples. We need to go from 10^p to 10^{2p}!
- With p=2, we start with 100 to 10,000
- This is the *Curse of Dimensionality* problem
- A clear need to reduce p (number of features).

Curse of Dimensionality

- The *Curse of Dimensionality* problem can be viewed from another point of view

- The number of features defined the hypothesis space we need to explore to find the best boundary.
- The more features, the larger the hypothesis space
- With *p* binary features, the number of possible combination of features is 2^p
- With a binary class feature,
  the hypothesis space is 2^{2^p}

Approaches

- The selection of the most appropriate features can be included in the model (e.g., 1R, decision tree) (embedded model)

- To select the attributes, we can
  - Rank them according to some criterion and then select the top *k* features
  - Select a minimum subset of features (subset selection)

Approaches

- With the subset selection (wrapper methods), we can apply
  - Forward selection: start with an empty set, and add one feature at the time
  - Backward selection: start with all features, then remove one at the time
  - Random:
  - Other variants exist
- Subset selection is also known in statistics
- When ranking features, we can consider different filter methods (used independently on each feature)
- For large scale problem, only filter methods

Focus on Text Classification

- Task: Given a document, predict a target class for it
  - Topical Text Classification: assign a topical label (e.g., sport, business, politics)
  - Authorship attribution: who wrote it
  - Genre: Is it an essay, a poem, a play, a fiction, …
  - Other: writing by a man or a woman?
- Text classification implies a lot of features
  - Words (e.g., armies), lemmas (e.g., army), punctuations
  - N-gram of characters, letters
  - POS, sequence of POS
  - Layout, spelling errors, …

Characteristics of Text Classification

- High dimensional feature space (in K, in G), larger than the number of training examples.
- Sparse document vectors: only a fraction of possible features appears in a given instance
- Heterogeneous use of terms: overlap between documents belonging to the same class is small. Many different formulations can be used to express the same idea).
- High level of redundancy: many features can be used to predict the correct class

Our Running Example

- Three target classes with 8 features

Feature	Red (50)	Green (75)	Blue (100)
color	50	75	100
red	48	2	4
green	2	73	3
blue	1	3	97
test	20	20	20
yellow	10	15	30
brown	25	2	50
sky	2	52	87

Example

- The *color* feature appears in all instances
  Not useful
- The *red*, *green*, *blue* features are more frequent in one class
  - As soon we observe a *red*, *green*, or *blue* feature, we have a good prediction of the target class
- The *brown* feature is rarely present in green target
- The *sky* feature is present in red or blue instances
- For the *test* and *yellow* feature … noisy?

Contingency Table

- Represent the situation with a contingency table
- Inspect the association between one feature t_k and the category c_j.

	Category c_j	**Category** $-c_j$	
Term t_k	*a*	*b*	*a + b*
Term $-t_k$	*c*	*d*	*c + d*
	a + c	*b + d*	*n = a+b+c+d*

Contingency Table

- Represent the situation with a contingency table
- Selecting the feature "green" and the class Green

	Category *Green*	**Category** *- Green*	
"green"	*73*	*5*	*78*
- "green"	*2*	*145*	*147*
	75	*150*	*n = 225*

Probability Estimates

- We need to estimate the probability of different events. We will consider the feature t_k and the target category c_j.
- Notation

  $Prob[c_j]$ Prob of having an instance of class c_j

  $Prob[t_k]$ Prob of having an instance with value t_k

  $Prob[t_k,c_j]$ Prob of having an instance of class c_j and with a feature t_k

  $Prob[-c_j]$ Prob of having an instance not belonging to class c_j

  $Prob[-t_k]$ Prob of having an instance not with the feature t_k

Probability Estimates

- Probability estimations

  $\text{Prob}[c_j] = (a+c) / n$

  $\text{Prob}[t_k] = (a+b) / n$

  $\text{Prob}[t_k,c_j] = a / n$

  For other cases, we can add these estimates

  $\text{Prob}[t_k|c_j] = a / (a+c)$

  $\text{Prob}[c_j|t_k] = a / (a+b)$

  $\text{Prob}[-t_k] = (c+d) / n$

  $\text{Prob}[-t_k,c_j] = c / n$

  $\text{Prob}[-t_k,-c_j] = d / n$

Contingency Table

$Prob[c_j]$	75 / 225 = 0.333
$Prob[t_k]$	78 / 225 = 0.346
$Prob[t_k,c_j]$	73 / 225 = 0.324
$Prob[t_k \mid c_j]$	73 / 75 = 0.973

	Category *Green*	**Category** **-** *Green*	
"green"	*73*	*5*	*78*
- *"green"*	*2*	*145*	*147*
	75	*150*	*n = 225*

Pointwise Mutual Information

- Compare the probability of having jointly the feature t_k and the target value c_j with the independence assumption

$$f(t_k, c_j) = log_2 \left[\frac{Prob[t_k, c_j]}{Prob[t_k] \cdot Prob[c_j]} \right]$$

$$= log_2 \left[Prob[t_k | c_j] \right] - log_2 \left[Prob[t_k] \right]$$

- No link between the feature and the target class, then the $PMI(t_k,c_j) \approx log(1) \approx 0$
- Positive association between the feature and the target class, then the $PMI(t_k,c_j) > 0$
- Negative association, $PMI(t_k,c_j) < 0$

Pointwise Mutual Information

- The presence of *red* in an example is a good evidence that the class is **red**.

	Category *Red*	**Category** -*Red*	
Term *red*	48	6	54
Term -*red*	2	169	171
	50	175	n = 225

$$PMI(red, \mathbf{red}) = log_2 \left[\frac{\frac{a}{n}}{\frac{a+b}{n} \cdot \frac{a+c}{n}} \right] = log_2 \left[\frac{\frac{48}{225}}{\frac{54}{225} \cdot \frac{50}{225}} \right] = 2$$

Pointwise Mutual Information

- The presence of *test* in an example does not provide any evidence for the class is **green**.

	Category *Green*	**Category** -*Green*	
Term *test*	20	40	60
Term -*test*	55	110	165
	75	150	n = 225

$$PMI(test, \mathbf{green}) = log_2 \left[\frac{\frac{a}{n}}{\frac{a+b}{n} \cdot \frac{a+c}{n}} \right] = log_2 \left[\frac{\frac{20}{225}}{\frac{60}{225} \cdot \frac{75}{225}} \right] = 0.0$$

Pointwise Mutual Information

- The presence of *blue* in an example is a good evidence that the class is not **red**.

	Category *red*	**Category** *-red*	
Term *blue*	1	100	101
Term *-blue*	49	75	124
	50	175	n = 225

$$PMI(blue, \mathbf{red}) = log_2 \left[\frac{\frac{a}{n}}{\frac{a+b}{n} \cdot \frac{a+c}{n}} \right] = log_2 \left[\frac{\frac{1}{225}}{\frac{101}{225} \cdot \frac{50}{225}} \right] = -4.489$$

Odds Ratio

- Compare the odd of having jointly the feature t_k and the target value c_j and the odd of having the feature t_k and not belonging to the class c_j

$$f(t_k, c_j) = \frac{\frac{Prob[t_k|c_j]}{1-Prob[t_k|c_j]}}{\frac{Prob[t_k|-c_j]}{1-Prob[t_k|-c_j]}} = \frac{Prob[t_k|c_j] \cdot (1 - Prob[t_k| - c_j])}{(1 - Prob[t_k|c_j]) \cdot Prob[t_k| - c_j]}$$

- No link between the feature and the target class, then the $OR(t_k,c_j) \approx 1$
- Positive association between the feature and the target class, then the $OR(t_k,c_j) > 1$
- Negative association, $OR(t_k,c_j) \approx 0$
- We can also have the log of the odds ratio.

Odds Ratio

- The presence of *red* in an example is a good evidence that the class is **red**.

	Category *red*	**Category** *-red*	
Term *red*	48	6	54
Term *-red*	2	169	171
	50	175	n = 225

$$OR(red, \mathbf{red}) = \frac{48/(48+2) \cdot (1 - (6/(6+169)))}{(1 - (48/(48+2))) \cdot 6/(6+169)} = 676.0$$

Odds Ratio

- The presence of *sky* in an example is an evidence that the class is not **red**.

	Category *red*	**Category** *-red*	
Term *sky*	2	139	141
Term *-sky*	48	36	84
	50	175	n = 225

$$OR(sky, \mathbf{red}) = \frac{2/(2+48) \cdot (1-(139/(139+36)))}{(1-(2/(2+48))) \cdot 139/(139+36)} = 0.01$$

Chi-Square

- Compare the joint distribution of having the feature t_k and the category c_j and distribution derived from independence

$$f(t_k, c_j) = \frac{n \cdot [(Prob[t_k, c_j] \cdot Prob[-t_k, -c_j]) - (Prob[t_k, -c_j] \cdot Prob[-t_k, c_j])]^2}{Prob[t_k] \cdot Prob[-t_k] \cdot Prob[c_j] \cdot Prob[-c_j]}$$

- No link between the feature and the target class, then the $CHI(t_k,c_j) \approx 0$ (or a small positive value)
- Positive (or negative) association between the feature and the target class, then the $CHI(t_k,c_j) >> 1$

Chi-Square

- The presence of *blue* in an example is a good evidence that the class is **blue**.

	Category *blue*	**Category** -*blue*	
Term *blue*	97	4	101
Term -*blue*	3	121	124
	100	125	n = 225

$$CHI(blue, \mathbf{blue}) = \frac{225 \cdot [(97/225 \cdot 121/225) – (4/225 \cdot 3/225)]^2}{101/225 \cdot 124/225 \cdot 100/225 \cdot 125/225} = 197.59$$

Chi-Square

- The presence of *yellow* in an example is a good evidence that the class is **red**? **NO**

	Category *red*	**Category** *-red*	
Term *yellow*	10	45	55
Term *-yellow*	40	130	170
	50	175	n = 225

$$CHI(yellow, \mathbf{red}) = \frac{225 \cdot [(10/225 \cdot 130/225) – (45/225 \cdot 40/225)]^2}{55/225 \cdot 170/225 \cdot 50/225 \cdot 175/225} = 0.688$$

Information Gain

- Compare the joint distribution of having the feature t_k and the target value c_j and a random distribution

$$f(t_k, c_j) = \sum_{c \in (c_j, -c_j)} \sum_{t \in (t_k, -t_k)} Prob[t, c] \cdot \log_2 \left[\frac{Prob[t, c]}{Prob[t] \cdot Prob[c]} \right]$$

- No link between the feature and the target class, then the $IG(t_k, c_j) \approx 0$ (or a small positive value)
- Positive association between the feature and the target class, then the $IG(t_k, c_j) > 0$
- Negative association between the feature and the target class, then the $IG(t_k, c_j) < 0$

Information Gain

- The presence of *red* in an example is a good evidence that the class is **red**.

	Category *red*	**Category** *-red*	
Term *red*	48	6	54
Term *-red*	2	169	171
	50	175	n = 225

$$IG(red, \mathbf{red}) = 48/225 \cdot \log_2 \left[\frac{48/225}{54/225 \cdot 50/225} \right] + 6/225 \cdot \log_2 \left[\frac{6/225}{54/225 \cdot 175/225} \right]$$
$$+2/225 \cdot \log_2 \left[\frac{2/225}{171/225 \cdot 50/225} \right] + 169/225 \cdot \log_2 \left[\frac{169/225}{171/225 \cdot 175/225} \right] = 0.148$$

Document Frequency

- The usefulness of a feature t_k for a given class c_j could be indicted by simpler function such as the document frequency (*df* or the number of instances where the feature in question does occur)
- Thus $f(t_k,c_j) = df(t_k,c_j)$
- The larger the value, the better the association between the feature and the class

- As a variant, we can compute the term frequency (*tf*)
- Thus $f(t_k,c_j) = tf(t_k,c_j)$

Back to our Example

- The starting data for three target classes and 8 features

Feature	Red (50)	Green (75)	Blue (100)
color	50	75	100
red	48	2	4
green	2	73	3
blue	1	3	97
test	20	20	20
yellow	10	15	30
brown	25	2	50
sky	2	52	87

Back to our Example

- Apply the *odds ratio* function, we obtain. Which features must we selected?

Feature	Red (50)	Green (75)	Blue (100)
color	1.00	1.00	1.00
red	676.00	0.05	0.06
green	0.05	1058.50	0.02
blue	0.02	0.02	978.08
test	2.25	1.00	0.53
yellow	0.72	0.69	1.71
brown	2.37	0.03	3.63
sky	0.01	1.55	8.80

Aggregation Operators

- We need to aggregate the values over the m classes

$$
\begin{aligned}
f_{max}(t_k) &= \max_{j=1}^{m} f(t_k, c_j) \\
f_{sum}(t_k) &= \sum_{j=1}^{m} f(t_k, c_j) \\
f_{wmean}(t_k) &= \sum_{j=1}^{m} Prob[c_j] \cdot f(t_k, c_j)
\end{aligned}
$$

Aggregation Operators

- Various studies tend to show that the max operator provide better result
- It seems better to select a feature giving good prediction for a given class than a feature performing, in mean, not so bad over all classes

Aggregation in our Example

- After applying the three aggregation operators on our OR data, we obtain

Feature	MAX	SUM	WMean
color	1.00	3.00	225.00
red	676.00	676.11	33,810.12
green	**1,058.50**	**1,058.57**	**79,392.28**
blue	**978.08**	**978.12**	**97,810.76**
test	2.25	3.78	240.63
yellow	1.71	3.12	259.10
brown	3.63	6.02	483.29
sky	8.80	10.36	996.67

Overall Result

- Various studies tend to show that the *max* operator provide better result
- Various papers in topical text categorization tends to show that:

  $(OR_{sum}\ ,\ NGL_{sum},\ GSS_{max}) > (X^2_{max},\ IG_{sum}) > (X^2_{wmean}) >> (PMI_{max},\ PMI_{wmean})$

F. Sebastiani. *Machine Learning in Automated Text Categorization*. ACM Computing Surveys, 34(1), 2002, 1-41

Conclusion

- Various text categorization tasks are possible
  binary or 2-class
  1-of-*m* (one label over a set of *m*)
  n-of-*m* (automatic indexing)
- Limited to select one ranking function
- Explore: using two (or more) functions in conjunction
- Interaction possible between the selection function and the classifier