MU5IN075
Network Analysis and Mining
13. Link prediction, a classification problem

Esteban Bautista-Ruiz, Lionel Tabourier

LIP6 – CNRS and Sorbonne Université

first_name.last_name@lip6.fr

January 4, 2022

Outline

A few examples

- recommendation on a social network
  People you may know on Facebook
- recommendation in general
  papers, news, contents on the web
- not only recommendation
  high throughput screening in drug discovery

Guessing a potential link from the current structure

The link prediction problem: temporal version

Problem description: temporal version

V is a fixed set of nodes,

- interactions known between t_0 and t'_0
- which links appear(/disappear) between t_1 and t'_1?

Liben-Nowell, Kleinberg - *JASIST, 2007*

graph observed during $[t_0, t'_0]$ $\rightarrow$ links appearing during $[t_1, t'_1]$

use features correlated with appearance of a link

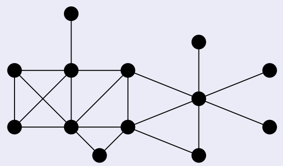
The link prediction problem: temporal version

Problem description: temporal version

V is a fixed set of nodes,

- interactions known between t_0 and t'_0
- which links appear(/disappear) between t_1 and t'_1?

Liben-Nowell, Kleinberg - *JASIST, 2007*



graph observed during $[t_0, t'_0] \rightarrow$ links appearing during $[t_1, t'_1]$

use features correlated with appearance of a link

Introduction – Context
A classification problem
Resolution examples

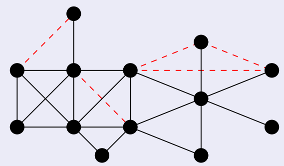
The link prediction problem: temporal version

Problem description: temporal version

V is a fixed set of nodes,

- interactions known between t0 and t'0
- which links appear(/disappear) between t1 and t'1?

Liben-Nowell, Kleinberg - JASIST, 2007



graph observed during [t0, t'0] → links appearing during [t1, t'1]

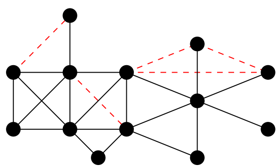
use features correlated with appearance of a link

4/26

The missing link problem

Principle

- suppose that the data crawling process missed links
- $\Rightarrow$ detect unseen links



Note for later

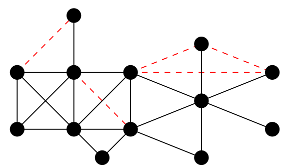
We consider large sparse graphs $\Rightarrow$ few edges for many pairs
probably difficult to predict with high accuracy...

Introduction – Context
A classification problem
Resolution examples

The missing link problem

Principle

- suppose that the data crawling process missed links
- ⇒ detect unseen links



Note for later

We consider large sparse graphs ⇒ few edges for many pairs
probably difficult to predict with high accuracy...

5/26

Outline

Classification problems

What is statistical classification?

- **classification**:
  fixed number of groups, a group for each data point
- **statistical**:
  based on comparison of the data point features to a population of already classified points

Classification in supervised learning

Reminder:

- prediction tasks using labeled data → supervised learning
- two main problems:
  - predicting a number (score, rating, measure,...): regression
  - predicting a category among a finite set: classification

Classification problems

What is statistical classification?

- **classification**:
  fixed number of groups, a group for each data point
- **statistical**:
  based on comparison of the data point features to a population of already classified points

Classification in supervised learning

Reminder:

- prediction tasks using labeled data → supervised learning
- two main problems:
  - predicting a number (score, rating, measure,...): regression
  - predicting a category among a finite set: classification

Some classic classification examples

Task	Classes	Features
species classification	species	shape, weight, ...
character recognition	a,b,c,...	shape, pixels
medical diagnosis	diseases	physical measurements
spam detection	spam / ham	words
link prediction	link / no link	network structure

The link prediction case

Remember our *note for later*...

- classify between two classes → binary
- in large graphs, many more unconnected pairs than edges
  ⇒ a class is much larger than the other → imbalanced

Some classic classification examples

Task	Classes	Features
species classification	species	shape, weight, ...
character recognition	a,b,c,...	shape, pixels
medical diagnosis	diseases	physical measurements
spam detection	spam / ham	words
link prediction	link / no link	network structure

The link prediction case

Remember our *note for later*...

- classify between two classes → binary
- in large graphs, many more unconnected pairs than edges
  ⇒ a class is much larger than the other → imbalanced

Introduction – Context
A classification problem
Resolution examples

Some classic classification examples

Task	Classes	Features
species classification	species	shape, weight, ...
character recognition	a,b,c,...	shape, pixels
medical diagnosis	diseases	physical measurements
spam detection	spam / ham	words
link prediction	link / no link	network structure

The link prediction case

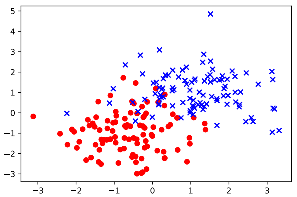
Remember our *note for later*...

- classify between two classes → binary
- in large graphs, many more unconnected pairs than edges
  ⇒ a class is much larger than the other → imbalanced

8/26

How to solve a classification problem

Example: binary classification, two features

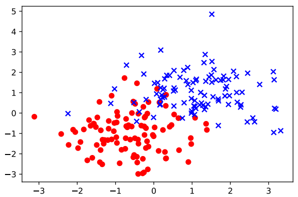


- how to draw frontiers?
- how to set parameters of a model?
- how to evaluate results?

Introduction – Context
A classification problem
Resolution examples

How to solve a classification problem

Example: binary classification, two features



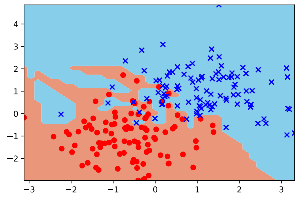
- how to draw frontiers?
- how to set parameters of a model?
- how to evaluate results?

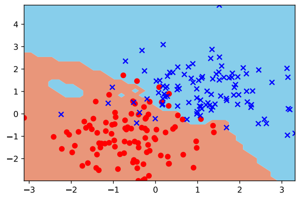
9/26

How to draw frontiers?

Select a model in the toolbox

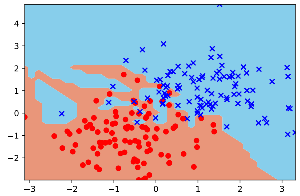
k-nearest neighbors
naive Bayes classifier
classification tree
support vector machine
neural networks
...





The problem of overfitting

(fr: surapprentissage)

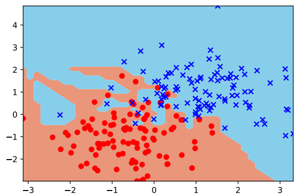


$\Rightarrow$ too particular to the training set

Evaluation must be done on a test set

The problem of overfitting

(fr: surapprentissage)

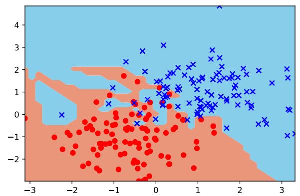


$\Rightarrow$ too particular to the training set

Evaluation must be done on a test set

The problem of overfitting

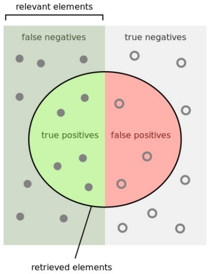
(fr: surapprentissage)



$\Rightarrow$ too particular to the training set

Evaluation must be done on a test set

How to evaluate results? (1)

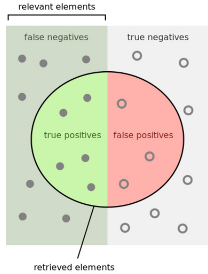


Confusion matrix

	prediction: +	prediction -
reality: +	true positive	false negative
reality: -	false positive	true negative

What do you think for link prediction?

How to evaluate results? (1)



Confusion matrix

	prediction: +	prediction -
reality: +	true positive	false negative
reality: -	false positive	true negative

What do you think for link prediction?

How to evaluate results? (2)

Standard metrics

- precision, **Pr** $= \frac{\#tp}{\#tp+\#fp}$
- recall, **Rc** $= \frac{\#tp}{\#tp+\#fn}$ also called sensitivity (*fr: rappel, sensibilité*)

nb: normalized metrics, think of extreme cases

Among many others. . .

- F-score $= \frac{2.\mathbf{Pr}.\mathbf{Rc}}{\mathbf{Pr}+\mathbf{Rc}}$ balance between precision and recall
- specificity (*fr: spécificité*), *ROC* curve,. . .

How to evaluate results? (2)

Standard metrics

- precision, **Pr** $= \frac{\#tp}{\#tp+\#fp}$
- recall, **Rc** $= \frac{\#tp}{\#tp+\#fn}$ also called sensitivity (*fr: rappel, sensibilité*)

nb: normalized metrics, think of extreme cases

Among many others. . .

- F-score $= \frac{2.\mathbf{Pr}.\mathbf{Rc}}{\mathbf{Pr}+\mathbf{Rc}}$ balance between precision and recall
- specificity (*fr: spécificité*), *ROC* curve,. . .

How to evaluate results? (3)

Misclassification importance depends on context

Spam detection: important not to class ham as spam
⇒ false positive ≫ false negative
⇒ favors precision over recall
Cancer diagnosis: capital not to miss a positive diagnosis
⇒ false negative ≫ false positive
⇒ favors recall over precision

Introduction – Context
A classification problem
Resolution examples

How to evaluate results? (3)

Misclassification importance depends on context

Spam detection: important not to class ham as spam
⇒ false positive ≫ false negative
⇒ favors precision over recall
Cancer diagnosis: capital not to miss a positive diagnosis
⇒ false negative ≫ false positive
⇒ favors recall over precision

14/26

Introduction – Context
A classification problem
Resolution examples

How to evaluate results? (3)

Misclassification importance depends on context

Spam detection: important not to class ham as spam
⇒ false positive ≫ false negative
⇒ favors precision over recall
Cancer diagnosis: capital not to miss a positive diagnosis
⇒ false negative ≫ false positive
⇒ favors recall over precision

14/26

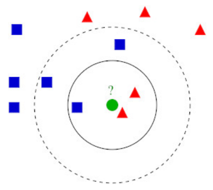
Introduction – Context
A classification problem
Resolution examples

Outline

K-nearest neighbors

Context

- Each data point is located in a space of features
- Each data point has a class *(ex: red triangles, blue squares)*



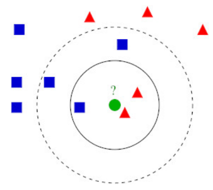
Principle

- For a new unlabeled data point *(ex: green circle)*: compute its distance to all labeled data points
- Prediction = dominant class among its *k* nearest neighbors

K-nearest neighbors

Context

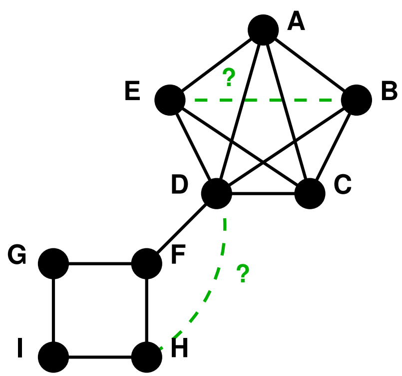
- Each data point is located in a space of features
- Each data point has a class *(ex: red triangles, blue squares)*



Principle

- For a new unlabeled data point *(ex: green circle)*: compute its distance to all labeled data points
- Prediction = dominant class among its *k* nearest neighbors

K-nearest neighbors: application to link prediction

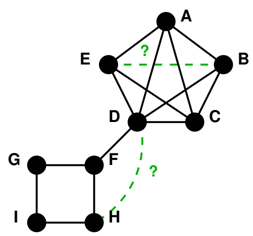


Prediction features

Structural characteristics

With $\mathcal{N}(i)$ the set of neighbors of node i

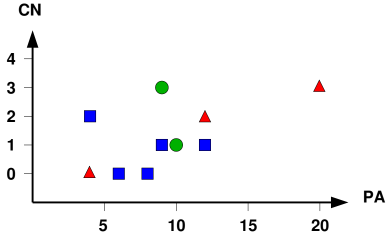
- number of common neighbors (CN) $|\mathcal{N}(i) \cap \mathcal{N}(j)|$
- preferential attachment index (PA) $|\mathcal{N}(i)|.|\mathcal{N}(j)|$



pair	CN	PA	edge or not
(A,B)	2	12	yes
(C,D)	3	20	yes
(C,F)	1	12	no
(B,F)	1	9	no
(H,A)	0	8	no
(H,G)	2	4	no
(I,G)	0	4	yes
(I,B)	0	6	no
(E,B)	3	9	?
(H,D)	1	10	?

Resolution

pair	CN	PA	edge or not
(A,B)	2	12	yes
(C,D)	3	20	yes
(C,F)	1	12	no
(B,F)	1	9	no
(H,A)	0	8	no
(H,G)	2	4	no
(I,G)	0	4	yes
(I,B)	0	6	no
(E,B)	3	9	?
(H,D)	1	10	?



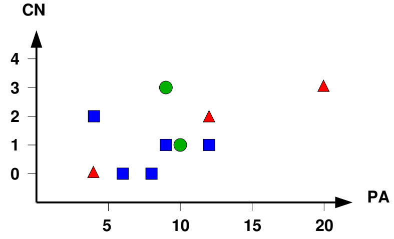
Features should be normalized before distance computation
reduced centered features: $x \rightarrow \frac{x-\bar{x}}{\sigma_x}$

On large graphs, computing all distances is inefficient

Does not work well for link prediction

Resolution

pair	CN	PA	edge or not
(A,B)	2	12	yes
(C,D)	3	20	yes
(C,F)	1	12	no
(B,F)	1	9	no
(H,A)	0	8	no
(H,G)	2	4	no
(I,G)	0	4	yes
(I,B)	0	6	no
(E,B)	3	9	?
(H,D)	1	10	?



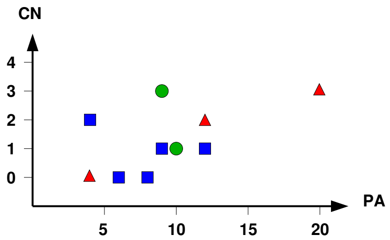
Features should be normalized before distance computation
reduced centered features: $x \rightarrow \frac{x-\bar{x}}{\sigma_x}$

On large graphs, computing all distances is inefficient

Does not work well for link prediction

Resolution

pair	CN	PA	edge or not
(A,B)	2	12	yes
(C,D)	3	20	yes
(C,F)	1	12	no
(B,F)	1	9	no
(H,A)	0	8	no
(H,G)	2	4	no
(I,G)	0	4	yes
(I,B)	0	6	no
(E,B)	3	9	?
(H,D)	1	10	?



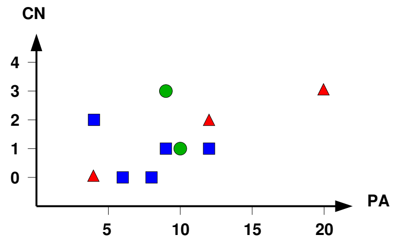
Features should be normalized before distance computation
reduced centered features: $x \rightarrow \frac{x-\bar{x}}{\sigma_x}$

On large graphs, computing all distances is inefficient

Does not work well for link prediction

Resolution

pair	CN	PA	edge or not
(A,B)	2	12	yes
(C,D)	3	20	yes
(C,F)	1	12	no
(B,F)	1	9	no
(H,A)	0	8	no
(H,G)	2	4	no
(I,G)	0	4	yes
(I,B)	0	6	no
(E,B)	3	9	?
(H,D)	1	10	?



Features should be normalized before distance computation
reduced centered features: $x \rightarrow \frac{x-\bar{x}}{\sigma_x}$

On large graphs, computing all distances is inefficient

Does not work well for link prediction

Classification using ranking

Principle

- Choose a feature, rank pairs with this feature
- Predict top T pairs according to this ranking

Advantages and drawbacks

- fast as we don't need to compute for all pairs in general
  ex: CN, ignore pairs of nodes at distance > 2
- only possible if higher score $\equiv$ more probable edge

Classification using ranking

Principle

- Choose a feature, rank pairs with this feature
- Predict top T pairs according to this ranking

Advantages and drawbacks

- fast as we don't need to compute for all pairs in general
  ex: CN, ignore pairs of nodes at distance > 2
- only possible if higher score $\equiv$ more probable edge

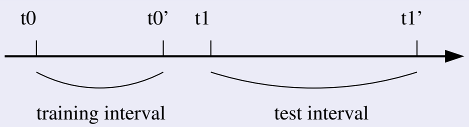
An example from the literature

Liben-Nowell, Kleinberg - *JASIST, 2007*

Datasets: scientific collaboration networks

- node = authors, link = co-publication
- publications in *DBLP, arXiv, Medline...*
- number of articles: a few thousands per year
- number of authors: a few thousands

Protocol



Year A to predict new collaborations in year $A+1$

Prediction features (part 1)

A closer look on local structure

- Number of common neighbors:

$$|\mathcal{N}(i) \cap \mathcal{N}(j)|$$

- Jaccard index:

$$\frac{|\mathcal{N}(i) \cap \mathcal{N}(j)|}{|\mathcal{N}(i) \cup \mathcal{N}(j)|}$$

- Adamic-Adar index:

$$\sum_{k \in \mathcal{N}(i) \cap \mathcal{N}(j)} \frac{1}{log(\delta(k))}$$

Prediction features (part 1)

A closer look on local structure

- Number of common neighbors:

$$|\mathcal{N}(i) \cap \mathcal{N}(j)|$$

- Jaccard index:

$$\frac{|\mathcal{N}(i) \cap \mathcal{N}(j)|}{|\mathcal{N}(i) \cup \mathcal{N}(j)|}$$

- Adamic-Adar index:

$$\sum_{k \in \mathcal{N}(i) \cap \mathcal{N}(j)} \frac{1}{log(\delta(k))}$$

Introduction – Context
A classification problem
Resolution examples

Prediction features (part 1)

A closer look on local structure

- Number of common neighbors:

$$|\mathcal{N}(i) \cap \mathcal{N}(j)|$$

- Jaccard index:

$$\frac{|\mathcal{N}(i) \cap \mathcal{N}(j)|}{|\mathcal{N}(i) \cup \mathcal{N}(j)|}$$

- Adamic-Adar index:

$$\sum_{k \in \mathcal{N}(i) \cap \mathcal{N}(j)} \frac{1}{log(\delta(k))}$$

22/26

Prediction features (part 2)

About other kinds of index

- preferential attachment index: $|\mathcal{N}(i)|.|\mathcal{N}(j)|$
  rely on the fact that high degree nodes tend to connect
- large-scale structure indices
  - *hitting time* from i to j:
    expected number of steps required for a random walk starting at i to reach j
    $\rightarrow$ rank by inverse hitting time

Prediction features (part 3)

Non-structural characteristics

- similarity indices between nodes i and j:
  - age
  - gender
  - for scientists: field of expertise
- and many other features $\rightarrow$ classification in general

Prediction features (part 3)

Non-structural characteristics

- similarity indices between nodes i and j:
  - age
  - gender
  - for scientists: field of expertise
- and many other features $\rightarrow$ classification in general

Quality assessment for link prediction

Prediction in large networks, **class imbalance problem**:
high risk of FP $\Rightarrow$ precision often low
we need a benchmark for comparison

A basic protocol

Liben-Nowell, Kleinberg - *JASIST, 2007*

- set N_{new}, the number of new links that appear
- keep the N_{new} top scoring items according to each feature
- compare to a random prediction

Introduction – Context
A classification problem
Resolution examples

Quality assessment for link prediction

Prediction in large networks, **class imbalance problem**:
high risk of FP $\Rightarrow$ precision often low
we need a benchmark for comparison

A basic protocol

Liben-Nowell, Kleinberg - *JASIST, 2007*

- set N_{new}, the number of new links that appear
- keep the N_{new} top scoring items according to each feature
- compare to a random prediction

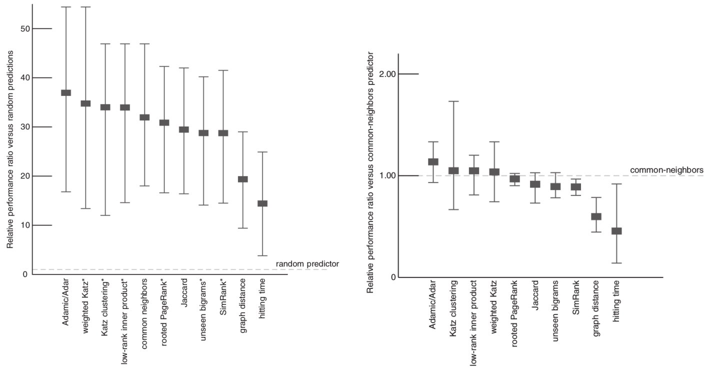
Introduction – Context
A classification problem
Resolution examples

Results

Predictor		astro-ph	cond-mat	gr-qc	hep-ph	hep-th
probability that a random prediction is correct		0.475%	0.147%	0.341%	0.207%	0.153%
graph distance (all distance-2 pairs)		*9.4*	*25.1*	*21.3*	*12.0*	*29.0*
common neighbors		**18.0**	**40.8**	**27.1**	**26.9**	**46.9**
preferential attachment		4.7	6.0	7.5	*15.2*	7.4
Adamic/Adar		*16.8*	**54.4**	**30.1**	**33.2**	**50.2**
Jaccard		*16.4*	**42.0**	19.8	**27.6**	*41.5*
SimRank	$\gamma = 0.8$	*14.5*	*39.0*	*22.7*	*26.0*	*41.5*
hitting time		6.4	23.7	*24.9*	3.8	13.3
hitting time—normed by stationary distribution		5.3	23.7	11.0	11.3	21.2

- probability that random prediction is correct is very low
- performance = factor improvement over random prediction

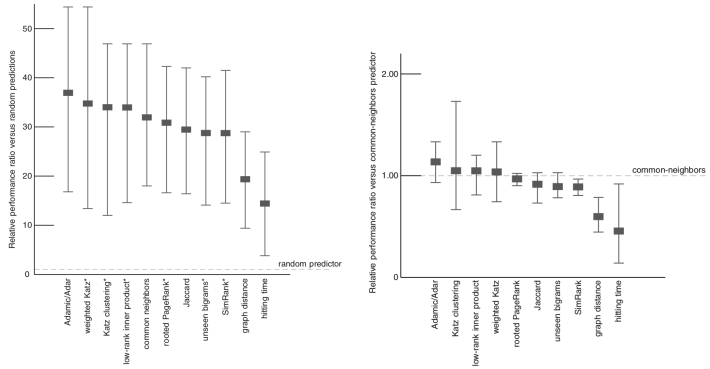
Results



Adamic-Adar usually efficient feature for link prediction

Introduction – Context
A classification problem
Resolution examples

Results



Adamic-Adar usually efficient feature for link prediction

26/26