THE UNIVERSITY OF MELBOURNE

Feature Selection

recap: ① PCA

linear comb of old features to new feature

② subset of feature

Semester 1, 2021

Ling Luo

Outline

- Feature selection methods
  - Wrappers
  - Embedded
  - Filtering
- Filtering methods
  - Pointwise Mutual Information (PMI)
  - Mutual Information (MI)
  - χ^2
- Common issues

Machine Learning

Outlook	Temp	Humidity	Windy	Play?
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
…	…	…	…	…

- How to do supervised machine learning?
  1) Pick a feature representation
  2) Compile data
  3) Pick a suitable model
  4) Train the model
  5) Classify validation/test data, evaluate results
  6) Go to Step 1

if feature violates some assumption model, then the evaluation result can be bad

Machine Learning

- Our tasks as Machine Learning experts:
  - Choose a model suitable for classifying the data according to the attributes
  - Choose useful attributes for classifying the data according to the model
    - ① Inspection? → histogram and distribution
    - ② Intuition?

problem:
① may not work for thousands of attributes

whether features are independent or linearly seperable

⇒ input and let model decides ⇒ can also be impacted by having irrelevant features

What are good features?

- Main goal:
  - Better performance according to some evaluation metric
- Side goals:
  - Seeing important features can suggest other important features
  - Fewer features → smaller models → faster answer
    - More accurate answer >> faster answer

if time complexity isn't a big issue

Methods

- Wrappers
- Embedded
- Filtering

① full wrapper: subset of attribute
② Greedy approach: sequential forward selection
③ Ablation approach: sequential backward selection

⇒ DT
regression with regularisation
LASSO L1
Ridge

MI
PMI
Chi 2

Wrappers

- Choose subset of attributes that give best performance on the validation data
- For example, for the weather data

Train model on:	**Evaluate**
{Outlook}	0.65
{Temperature}	0.6
…	…
{Outlook, Temperature}	0.75
…	…
{Outlook, Temperature, Humidity}	0.8
…	…

Pick the best feature set

Wrappers

- Advantage:
  - Can find the feature set with optimal performance on validation data for this learner
- Disadvantage:
  - Not practical, takes a long time

brute force approach

Wrappers

all combinations

- How long does the full wrapper method take?
  - Assume we have a fast method (e.g. Naïve Bayes) over a data set of medium size (~50K instances)
  - If each train-evaluate cycle takes 10 seconds to complete,

    $\frac{1}{6}$ min
  - For *m* attributes

    → empty set
    - $(2^m - 1)$ combinations → $\approx \frac{2^m}{6}$ minutes
    - $m = 10$ → ≈ 3 hours
    - $m = 60$ → $\approx 3.2^{15}$ hours
  - Only practical for very small data sets

More Practical Wrappers

- **Greedy Approach: sequential forward selection**
  - Train and evaluate model on each single attribute
  - Choose the best attribute eg A
  - Until convergence: conside {A, ?} where ? belongs to the remaining set
    - Train and evaluate model on best attribute(s), plus each remaining single attribute
    - Choose best attribute out of the remaining set
  - Termination condition: performance (e.g. accuracy) stops increasing

measurement metric: final prediction accuracy

More Practical Wrappers

- **Greedy Approach: sequential forward selection**
  - Running time: takes $\frac{m(m+1)}{2}$ $(\rightarrow = m + (m-1) + \cdots + 1)$ cycles for m attributes $O(m^2)$
  - In practice, converges much more quickly than this
  - Can convergence to a sub-optimal (or even bad) solution
  - Assumes independence of attributes

More Practical Wrappers

- **Ablation Approach: sequential backward selection**
  - Start with all attributes
  - Remove one attribute, train and evaluate model
  - Until divergence:
    - From remaining attributes, remove each attribute, train and evaluate model
    - Remove attribute that causes least performance degradation

      most likely redundant feature
  - Termination condition: performance (e.g. accuracy) starts to degrade by more than threshold ε

More Practical Wrappers

- **Ablation Approach: sequential backward selection**
  - Advantages:
    - Removes most of irrelevant attributes at the start
    - Performs best when the optimal subset is large
  - Disadvantages:

    $O(m^2)$ early iteration with more attributes can be slower

    - Running time: cycles can be slower with more attributes
    - Not feasible on large data sets

Embedded

- Embedded methods: in-built feature selection
  Models perform feature selection as *part of the algorithm*, for example:
  - Decision trees
  - Regression model with regularisation, e.g. linear regression with L1-norm regularisation (LASSO) (more about this later)
- Still benefit from other feature selection approaches

Filtering Methods

don't need to train model
don't rely on model prediction

- Intuition: evaluate "goodness" of each attribute
- Most popular strategy
- Consider each attribute separately: linear time in number of attributes

eg. decision tree
use gain ratio
(has been embedded to decision tree)

- What makes a single feature good?
  - Well correlated with interesting class

(correlation)

Good Features?

a_1	a_2	c
Y	Y	Y
Y	N	Y
N	Y	N
N	N	N

Which attribute, $\mathbf{a_1}$ or $\mathbf{a_2}$, is good?

a1

Good Features?

a_1	a_2	c
Y	Y	Y
Y	N	Y
N	Y	N
N	N	N

a_1 is probably good

Good Features?

a_1	a_2	c
Y	Y	Y
Y	N	Y
N	Y	N
N	N	N

a_2 is probably not good

Filtering Methods

- Pointwise Mutual Information (PMI) dependency of two variable
- Mutual Information (MI)
- χ^2

Pointwise Mutual Information PMI.

- Independence: the following formula holds if attribute *A* is independent from class *C*

$$P(A,C) = P(A)P(C) \qquad P(C|A) = P(C)$$

⇒ A & C are independent

$P(C|A) \gg P(C)$ → positively correlated

- If $\frac{P(A,C)}{P(A)P(C)} \gg 1$, attribute and class occur together much more often than randomly.

positively correlated

- If $\frac{P(A,C)}{P(A)P(C)} \approx 1$, attribute and class are independent, and they occur together as often as we would expect from random chance.
- If $\frac{P(A,C)}{P(A)P(C)} \ll 1$, attribute and class are negatively correlated.

$P(C|A) \ll P(C)$ → negatively correlated.

Pointwise Mutual Information

- Pointwise Mutual Information

check feature & class pair check positive / 0 / negative

$$PMI(A = a, C = c) = \log_2 \frac{P(a,c)}{P(a)P(c)}$$

- Best attributes: most correlated with class, the attributes with greatest PMI

PMI Example

a_1	a_2	c
Y	Y	Y
Y	N	Y
N	Y	N
N	N	N

$P(a_1)$ means $P(a_1 = Y)$, Y is the "interesting" value of a binary attribute

P(c=Y)

$$P(a_1) = \frac{2}{4}, P(c) = \frac{2}{4}, P(a_1, c) = \frac{2}{4}$$

$$PMI(a_1, c) = \log_2 \frac{\frac{1}{2}}{\frac{1}{2} \cdot \frac{1}{2}} = \log_2 2 = 1$$

PMI Example

a_1	a_2	c
Y	Y	Y
Y	N	Y
N	Y	N
N	N	N

$$P(a_2) = \frac{2}{4}, P(c) = \frac{2}{4}, P(a_2, c) = \frac{1}{4}$$

$$PMI(a_2, c) = \log_2 \frac{\frac{1}{4}}{\frac{1}{2} \cdot \frac{1}{2}} = \log_2 1 = 0$$

对于c=Y

PMI(a1=Y, c=Y)

>PMI(a2=Y, c=Y)

a1 is better than a2

a_1 is better than a_2

Find Good Features

Summary: What makes a single feature good?

① a and c

- Well correlated with interesting class
  - Knowing a lets us predict c with more confidence

② $\bar{a}$ and c

- Reverse correlated with interesting class
  - Knowing $\bar{a}$ (not a) lets us predict c with more confidence

③ a and $\bar{c}$

- Well correlated or reverse correlated with uninteresting class
  - Knowing a lets us predict $\bar{c}$ with more confidence
  - Usually not quite as good, but still useful

④ reverse correlated with uninteresting class

$\bar{a}$ and $\bar{c}$

Mutual Information

use PMI

use entropy $I(X,Y)=H(X|Y)-H(X)$
$=H(Y|X)-H(Y)$

$I(X,Y)\in[0,+\infty)$

- Mutual Information: consider the PMIs of all the combinations of $a, \bar{a}$ and $c, \bar{c}$

PMI

$$MI(A,C) = P(a,c)\boxed{\log_2 \frac{P(a,c)}{P(a)P(c)}} + P(\bar{a},c)\log_2 \frac{P(\bar{a},c)}{P(\bar{a})P(c)} +$$
$$P(a,\bar{c})\log_2 \frac{P(a,\bar{c})}{P(a)P(\bar{c})} + P(\bar{a},\bar{c})\log_2 \frac{P(\bar{a},\bar{c})}{P(\bar{a})P(\bar{c})}$$

- Often written more compactly as:

binary.

$$MI(A,C) = \sum_{i\in\{a,\bar{a}\}} \sum_{j\in\{c,\bar{c}\}} P(i,j)\log_2 \frac{P(i,j)}{P(i)P(j)}$$

- $0 \log_2 0$ is defined as 0

Contingency Tables

- Compact representation of these frequency counts

	$a = Y$	$a = N, (\bar{a})$	Total
$c = Y$	$\sigma(a, c)$	$\sigma(\bar{a}, c)$	$\sigma(c)$
$c = N, (\bar{c})$	$\sigma(a, \bar{c})$	$\sigma(\bar{a}, \bar{c})$	$\sigma(\bar{c})$
Total	$\sigma(a)$	$\sigma(\bar{a})$	M

- Compute $P(a, c), P(a), P(c)$ etc. based on the table

$$P(a, c) = \frac{\sigma(a, c)}{M}$$

Contingency Tables

- Contingency Tables for toy example with attributes a_1 and a_2

a_1	$a = Y$	$a = N$	Total
$c = Y$	2	0	2
$c = N$	0	2	2
Total	2	2	4

a_2	$a = Y$	$a = N$	Total
$c = Y$	1	1	2
$c = N$	1	1	2
Total	2	2	4

Mutual Information Example

- Contingency Tables for toy example: attribute a_1

a_1	$a = Y$	$a = N$	Total
$c = Y$	2	0	2
$c = N$	0	2	2
Total	2	2	4

$$P(a_1) = \frac{2}{4}, P(c) = \frac{2}{4}, P(\overline{a_1}) = \frac{2}{4}, P(\bar{c}) = \frac{2}{4}$$

$$P(a_1, c) = \frac{2}{4}, P(\overline{a_1}, c) = 0, P(a_1, \bar{c}) = 0, P(\overline{a_1}, \bar{c}) = \frac{2}{4}$$

Mutual Information Example

- MI for a_1

$$MI(A,C) = P(a_1,c)\log_2\frac{P(a_1,c)}{P(a_1)P(c)} + P(\overline{a_1},c)\log_2\frac{P(\overline{a_1},c)}{P(\overline{a_1})P(c)} +$$

$$P(a_1,\bar{c})\log_2\frac{P(a_1,\bar{c})}{P(a_1)P(\bar{c})} + P(\overline{a_1},\bar{c})\log_2\frac{P(\overline{a_1},\bar{c})}{P(\overline{a_1})P(\bar{c})}$$

$$= \frac{1}{2}\log_2\frac{\frac{1}{2}}{\frac{1}{2}\cdot\frac{1}{2}} + 0\log_2\frac{0}{\frac{1}{2}\cdot\frac{1}{2}} + 0\log_2\frac{0}{\frac{1}{2}\cdot\frac{1}{2}} + \frac{1}{2}\log_2\frac{\frac{1}{2}}{\frac{1}{2}\cdot\frac{1}{2}}$$

$$= \frac{1}{2}\cdot 1 + 0 + 0 + \frac{1}{2}\cdot 1 = 1$$

Mutual Information Example

- Contingency Tables for toy example: attribute a_2

a_2	$a = Y$	$a = N$	Total
$c = Y$	1	1	2
$c = N$	1	1	2
Total	2	2	4

$$P(a_2) = \frac{2}{4}, P(c) = \frac{2}{4}, P(\overline{a_2}) = \frac{2}{4}, P(\bar{c}) = \frac{2}{4}$$

$$P(a_2, c) = \frac{1}{4}, P(\overline{a_2}, c) = \frac{1}{4}, P(a_2, \bar{c}) = \frac{1}{4}, P(\overline{a_2}, \bar{c}) = \frac{1}{4}$$

Mutual Information Example

- MI for a_2

$$MI(A,C) = P(a_2,c)\log_2 \frac{P(a_2,c)}{P(a_2)P(c)} + P(\overline{a_2},c)\log_2 \frac{P(\overline{a_2},c)}{P(\overline{a_2})P(c)} +$$

$$P(a_2,\bar{c})\log_2 \frac{P(a_2,\bar{c})}{P(a_2)P(\bar{c})} + P(\overline{a_2},\bar{c})\log_2 \frac{P(\overline{a_2},\bar{c})}{P(\overline{a_2})P(\bar{c})}$$

$$= \frac{1}{4}\log_2 \frac{\frac{1}{4}}{\frac{1}{2}\cdot\frac{1}{2}} + \frac{1}{4}\log_2 \frac{\frac{1}{4}}{\frac{1}{2}\cdot\frac{1}{2}} + \frac{1}{4}\log_2 \frac{\frac{1}{4}}{\frac{1}{2}\cdot\frac{1}{2}} + \frac{1}{4}\log_2 \frac{\frac{1}{4}}{\frac{1}{2}\cdot\frac{1}{2}}$$

$$= 4\cdot\frac{1}{4}\cdot 0 = 0$$

a_1 is better than a_2

Chi-Square

- Similar idea with MI, but different solution
- Conduct statistical test to check the independence of a feature and the class
- Contingency table

W, X, Y, Z are count

	$a = Y$	$a = N, (\bar{a})$	Total
$c = Y$	W	X	$W + X$
$c = N, (\bar{c})$	Y	Z	$Y + Z$
Total	$W + Y$	$X + Z$	M

Chi-Square

- If a and c were independent, what value would we expect to be in W?
- Independence → $P(a,c) = P(a)P(c)$

$$\frac{\sigma(a,c)}{M} = \frac{\sigma(a)}{M} \cdot \frac{\sigma(c)}{M}$$

$$\sigma(a,c) = \frac{\sigma(a)\sigma(c)}{M}$$

$$E(W) = \frac{(W+Y)(W+X)}{W+X+Y+Z}$$

Chi-Square

- Compare the value actually observed $O(W)$ with the expected value $E(W)$ ($W = \sigma(a, c)$)
  - $O(W) \gg E(W)$: *a* occurs more often with *c* than we would expect at random – **predictive** positively correlated
  - $O(W) \ll E(W)$: *a* occurs less often with *c* than we would expect at random – **predictive** negatively correlated
  - $O(W) \approx E(W)$: *a* occurs as often with *c* as we would expect at random – **not predictive**
- Similarly with X, Y, Z

Chi-Square

- Calculation

$$\chi^2 = \frac{(O(W) - E(W))^2}{E(W)} + \frac{(O(X) - E(X))^2}{E(X)} + \frac{(O(Y) - E(Y))^2}{E(Y)} + \frac{(O(Z) - E(Z))^2}{E(Z)}$$

$$\boxed{\chi^2 = \sum_{i \in \{a,\bar{a}\}} \sum_{j \in \{c,\bar{c}\}} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}}$$

- Fit χ^2 to a chi-square distribution
- χ^2 becomes much greater when $|O - E|$ is large but E is small
- **High value of χ^2 indicates the dependency** between a feature and the class.

Chi-Square Example

- Contingency Tables for toy example attribute a_1

Observed values

a_1	$a = Y$	$a = N$	Total
$c = Y$	2	0	2
$c = N$	0	2	2
Total	2	2	4

Expected values (independent)

a_1	$a = Y$	$a = N$	Total
$c = Y$	1	1	2
$c = N$	1	1	2
Total	2	2	4

Chi-Square Example

- χ^2 for a_1

$$\chi^2 = \frac{(O_{a,c} - E_{a,c})^2}{E_{a,c}} + \frac{(O_{\bar{a},c} - E_{\bar{a},c})^2}{E_{\bar{a},c}} + \frac{(O_{a,\bar{c}} - E_{a,\bar{c}})^2}{E_{a,\bar{c}}} + \frac{(O_{\bar{a},\bar{c}} - E_{\bar{a},\bar{c}})^2}{E_{\bar{a},\bar{c}}}$$

$$= \frac{(2-1)^2}{1} + \frac{(0-1)^2}{1} + \frac{(0-1)^2}{1} + \frac{(2-1)^2}{1}$$

$$= 4$$

Chi-Square Example

- Contingency Tables for toy example attribute a_2

Observed values

a_2	$a = Y$	$a = N$	Total
$c = Y$	1	1	2
$c = N$	1	1	2
Total	2	2	4

Expected values (independent)

a_2	$a = Y$	$a = N$	Total
$c = Y$	1	1	2
$c = N$	1	1	2
Total	2	2	4

Chi-Square Example

- χ^2 for a_2

$$\chi^2 = \frac{(O_{a,c} - E_{a,c})^2}{E_{a,c}} + \frac{(O_{\bar{a},c} - E_{\bar{a},c})^2}{E_{\bar{a},c}} + \frac{(O_{a,\bar{c}} - E_{a,\bar{c}})^2}{E_{a,\bar{c}}} + \frac{(O_{\bar{a},\bar{c}} - E_{\bar{a},\bar{c}})^2}{E_{\bar{a},\bar{c}}}$$

$$= 0$$

- All observed values are equal to expected values
- Higher χ^2 indicates dependency, so a_1 is more predictive than a_2

Common Issues

Types of Attributes

Nominal attributes of multiple values → one-hot encoding

Outlook = {sunny, overcast, rainy}

sunny = [1, 0, 0]
overcast = [0, 1, 0]
rainy = [0, 0, 1]

- **Strategy 1**: Treat as multiple binary attributes
  - Convert to three features

    Outlook = sunny → sunny = Y, overcast = N, rainy = N
  - Use measures as given
  - But results can be difficult to interpret regarding the original feature

    For example, Outlook=sunny is useful, but Outlook=overcast and Outlook=rainy are not useful... Should we use Outlook?

Types of Attributes

- **Strategy 2**: Expand formulae and contingency tables

$Outlook$	$Sunny$	$Overcast$	$Rainy$	Total
$c = Y$	U	V	W	$U + V + W$
$c = N$	X	Y	Z	$X + Y + Z$
Total	$U + X$	$V + Y$	$W + Z$	M

$$MI(Outlook, C) = \sum_{i \in \{s,o,r\}} \sum_{j \in \{c,\bar{c}\}} P(i,j) \log_2 \frac{P(i,j)}{P(i)P(j)}$$

# of comb = # of attributes * # of outcomes

Types of Attributes

Continuous Attributes

1. - Estimate probabilities $P(a, c), P(a), P(c)$ etc. by fitting a distribution such as Gaussian
2. - Discretise values

Multi-class Problems

- Multiclass classification tasks are usually much more difficult than binary classification task
- For example, predict geotag for Melbourne, Sydney, Brisbane, Perth and Adelaide based on words in a post
  - How about these features: swanston, fed, mcg, docklands, afl? → good to distinguish Mel from other cities but they may not be predictive enough for other classes
  - Need to make a point of selecting features for each class to give our classifier the best chance of predicting every class correctly. (not just predict one class)

SVM
1-vs-rest
or all pairs

Summary

- Feature selection methods
  - Wrappers, embedded and filtering
- Popular filters: PMI, MI and χ^2
  - How to use them? What are the results going to look like?
- Importance of feature selection
  - necessary for distance-based models, e.g. kNN curse of dimensionality
  - Naive Bayes/Decision Trees, to a lesser extent
  - SVMs can work well without feature selection

work well in high dimension

References

- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. Introduction to Data Mining. Pearson, 2018.
- Ian Witten, Eibe Frank, and Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 3rd edition, 2011.
- Isabelle Guyon, and Andre Elisseeff. 2003. An introduction to variable and feature selection. The Journal of Machine Learning Research. Vol3, 1157–1182
- Yiming Yang, Jan Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of the Fourteenth International Conference on Machine Learning, 412–420

accurate model { uncorrelate between features ⇒ more info
high correlation between feature and label

Question 2

1 / 1 pts

Using full Wrapper method, the best model with 4 features always contains the set of features involved in the best model with 3 features.

- True
- Correct! False

Using full Wrapper method, it compares all combinations of features, and the best model with 4 features may not contain the set of features involved in the best model with 3 features. For example, two features working well together may be selected in 4-feature model, which replaces a certain feature in the best model with 3 features. However, for greedy approach wrapper, this statement is true, because greedy approach adds feature incrementally.

Question 3

0 / 1 pts

Which one of the following choices is an ideal value for PMI between the feature and the class?

- $+\varepsilon$ (a small positive value)
- 0
- Correct Answer $+\infty$
- You Answered 1

PMI is defined as log (P(a, b)/P(a)P(b)), which can be converted to log (P(a|b)/P(a)). In the ideal case, the input and the output are highly correlated, so that P(a|b) ≈1, and when P(a) is a small proabiltity value, in theory PMI can go to infinity.

PCA → feature reduction approach rather than feature selection