Machine Learning W9 Tutorial

COMP30027 | Sandy Luo

Overview

Feature Selection

Concept, code

Model Evaluation

General

Feature Selection

Q1:

Given the following dataset, we wish to perform feature selection, where the class to predict is PLAY:

ID	Outlook	Temp	Humid	Wind	PLAY
A	S	H	H	F	N
B	S	H	H	T	N
C	O	H	H	F	Y
D	R	M	H	F	Y
E	R	C	N	F	Y
F	R	C	N	T	N

1. Which of $Humid = H$ and $Wind = T$ has the greatest *pointwise mutual information* with the class Y? What about class N?
2. Which of the attributes has the greatest mutual information for the PLAY class as a whole?

Q1(a):

Given the following dataset, we wish to perform feature selection, where the class to predict is PLAY:

ID	Outlook	Temp	Humid	Wind	PLAY
A	S	H	H	F	N
B	S	H	H	T	N
C	O	H	H	F	Y
D	R	M	H	F	Y
E	R	C	N	F	Y
F	R	C	N	T	N

1. Which of $Humid = H$ and $Wind = T$ has the greatest *pointwise mutual information* with the class Y? What about class N?

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

Q1(a):

- P(Humid=H) = 4/6
- P(Wind=T) = 2/6
- P(C=Y) = 3/6
- P(C=N) = 3/6

Given the following dataset, we wish to perform feature selection, where the class to predict is PLAY:

ID	Outlook	Temp	Humid	Wind	PLAY
A	S	H	H	F	N
B	S	H	H	T	N
C	O	H	H	F	Y
D	R	M	H	F	Y
E	R	C	N	F	Y
F	R	C	N	T	N

- P(Humid=H, C=Y) = 2/6
- P(Humid=H, C=N) = 2/6
- P(Wind=T, C=Y) = 0
- P(Wind = T, C=N) = 2/6

1. Which of $Humid = H$ and $Wind = T$ has the greatest *pointwise mutual information* with the class Y? What about class N?

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

Q1(a):

- P(Humid=H) = 4/6
- P(Wind=T) = 2/6
- P(C=Y) = 3/6
- P(C=N) = 3/6
- P(Humid=H, C=Y) = 2/6
- P(Humid=H, C=N) = 2/6
- P(Wind=T, C=Y) = 0
- P(Wind = T, C=N) = 2/6

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

$$PMI(Humid=H, PLAY=Y) = log_2\frac{P(Humid=H \cap PLAY=Y)}{P(Humid=H)P(PLAY=Y)} = log_2\frac{(2/6)}{(4/6)(3/6)} = log_2(1) = 0$$

- PMI = 0 → Humid is (perfectly) uncorrelated with play = Y

$$PMI(Wind=T, PLAY=Y) = log_2\frac{P(Wind=T \cap PLAY=Y)}{P(Wind=T)P(PLAY=Y)} = log_2\frac{(0/6)}{(2/6)(3/6)} = log_2(0) = -\infty$$

- PMI = -\inf → Wind is (perfectly) negatively correlated with play = Y

Q1(a):

- P(Humid=H) = 4/6
- P(Wind=T) = 2/6
- P(C=Y) = 3/6
- P(C=N) = 3/6
- P(Humid=H, C=Y) = 2/6
- P(Humid=H, C=N) = 2/6
- P(Wind=T, C=Y) = 0
- P(Wind = T, C=N) = 2/6

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

PMI(Humid=H, C=N) = log_2(P(Humid=H, C=N)/P(Humid=H)P(C=N))

= log_2((2/6)/(4/6)(3/6)) = log_2(1) = 0

- PMI = 0 → Humid is (perfectly) uncorrelated with play = N

PMI(Wind=T, C=N) = log_2(P(Wind=T, C=N)/P(Wind=T)P(C=N)) =

log_2((2/6)/(2/6)(3/6)) = log_2(2) = 1

- PMI = 1 → Wind is positively correlated with play = N

Q1(b):

Given the following dataset, we wish to perform feature selection, where the class to predict is PLAY:

ID	Outlook	Temp	Humid	Wind	PLAY
A	S	H	H	F	N
B	S	H	H	T	N
C	O	H	H	F	Y
D	R	M	H	F	Y
E	R	C	N	F	Y
F	R	C	N	T	N

Consider the PMI of every possible attribute value−class combination, weighted by the proportion of instances that actually had that combination

Which of the attributes has the greatest mutual information for the PLAY class as a whole?

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

$$MI(X, C) = \sum_{x \in X} \sum_{c \in \{Y,N\}} P(c, x) PMI(x, c)$$

Q1(b):

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

$$MI(X,C) = \sum_{x \in X} \sum_{c \in \{Y,N\}} P(c,x)PMI(x,c)$$

$$MI(Outlook) = P(S,Y)PMI(S,Y) + P(O,Y)PMI(O,Y) + P(R,Y)PMI(R,Y) + P(S,N)PMI(S,N) + P(O,N)PMI(O,N) + P(R,N)PMI(R,N)$$
$$= \frac{0}{6}log_2\frac{(0/6)}{(2/6)(3/6)} + \frac{1}{6}log_2\frac{(1/6)}{(1/6)(3/6)} + \frac{2}{6}log_2\frac{(2/6)}{(3/6)(3/6)} + \frac{2}{6}log_2\frac{(2/6)}{(2/6)(3/6)} + \frac{0}{6}log_2\frac{(0/6)}{(1/6)(3/6)} + \frac{1}{6}log_2\frac{(1/6)}{(3/6)(3/6)}$$
$$= 0 + (0.1667)(1) + (0.3333)(0.4150) + (0.3333)(1) + 0 + (0.1667)(-0.5850)$$
$$= 0.541$$

$$MI(Temp) = P(H,Y)PMI(H,Y) + P(M,Y)PMI(M,Y) + P(C,Y)PMI(C,Y) + P(H,N)PMI(H,N) + P(M,N)PMI(M,N) + P(C,N)PMI(C,N)$$
$$= \frac{1}{6}log_2\frac{(1/6)}{(3/6)(3/6)} + \frac{1}{6}log_2\frac{(1/6)}{(1/6)(3/6)} + \frac{1}{6}log_2\frac{(1/6)}{(2/6)(3/6)} + \frac{2}{6}log_2\frac{(2/6)}{(3/6)(3/6)} + \frac{0}{6}log_2\frac{(0/6)}{(1/6)(3/6)} + \frac{1}{6}log_2\frac{(1/6)}{(2/6)(3/6)}$$
$$= (0.1667)(-0.5850) + (0.1667)(1) + (0.1667)(0) + (0.3333)(0.4150) + 0 + 0$$
$$= 0.208$$

- MI(Humid) = 0, MI(Wind) = 0.459

Model Evaluation

Q2(a):

What is the difference between model <u>bias</u> and model <u>variance</u>?

Q2(a):

What is the difference between model <u>bias</u> and model <u>variance</u>?

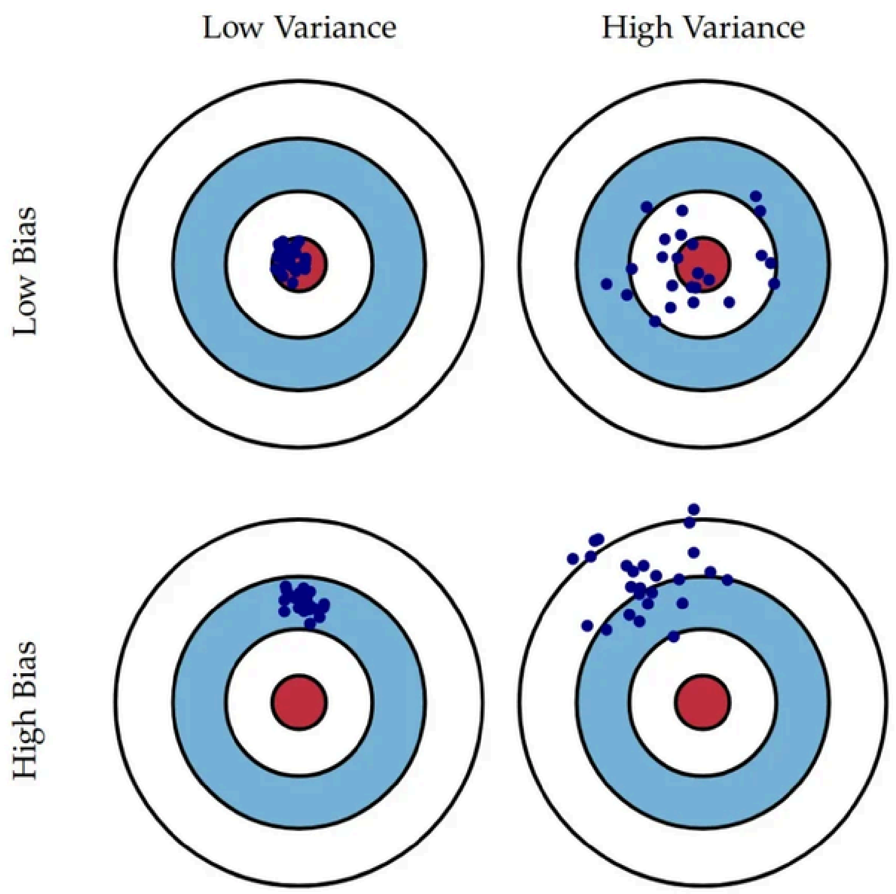

Fig. 1 Graphical illustration of bias and variance.

- **Bias:**
  - ***Systematically* producing similar errors**
  - e.g. predicted class *distribution* != actual
  - e.g. bias towards majority class
- **Variance**:
  - **Measure of inconsistency**
  - Difference in classifications w/ diff. training sets from the same population

Q2(b):

Describe the behaviour of a classifier with high bias and low variance.

Q2(b):

Describe the behaviour of a classifier with high bias and low variance.

- **High** bias → **systematically** product **similar errors**
- **Low** variance → **little** random error

- Therefore, this model will consistently produce the same type of wrong predictions

Q2(c):

Describe the behaviour of a classifier with low bias and high variance

Q2(c):

Describe the behaviour of a classifier with low bias and high variance

- **Low** bias → **DON'T** systematically product similar errors
- **High** variance → **A LOT OF** random errors

- Model will make a lot of inconsistent, random errors
  - Different types of errors
  - Error rate might be low on one set of data but high on another
- Distribution of predictions should match the true distribution (unbiased) but which instances are assigned to which labels may be quite variable.

Q3(a):

Explain how these strategies help reduce model overfitting:

- **Use of a validation set (e.g., cross-validation)**

Q3(a):

Explain how these strategies help reduce model overfitting:

- **Use of a validation set (e.g., cross-validation)**

- Only determining performance on training data is prone to overfitting
- Want independent (unseen) data to compare performance b/w models

- Trade-off b/w size of training & validation set
  - Can use cross-validation
  - Time trade-off

Q3(b):

Explain how these strategies help reduce model overfitting:

- **Model ensembling (e.g., random forests)**

Q3(b):

Explain how these strategies help reduce <u>model overfitting</u>:

- **Model ensembling (e.g., random forests)**

*Z_i...Z_n i.i.d.

- From statistics, averaging reduces variance

$$Var(\frac{1}{N}\sum_i Z_i) = \frac{1}{N}Var(Z_i)$$

- Average several models → decrease overall variance
- Technically doesn't introduce more bias

Q4:

Explain the difference between ***evaluation*** bias and ***model*** bias.

Q4:

Explain the difference between *evaluation* bias and *model* bias.

- **Model** bias:
  - Systematic error in **modelling** process → unable to capture the true relationship between features and the target
    - Leads to underfitting (high bias = oversimplified model).
  - Causes: Model architecture, feature selection, assumptions
- **Evaluation** bias:
  - Systematic error in **evaluation** process → consistently over/under-estimating the true performance of the model
  - Causes: Evaluation metric, sampling bias, evaluation assumptions

Q5(a):

During training process, your model shows significantly different performance across different training sets.

- ***W*hat can be the reason?**

Q5(a):

During training process, your model shows significantly different performance across different training sets.

- ***What can be the reason?***

- Large variety (change) in predictions with small changes in data set
    - High variance → overfitting

Q5(b):

During training process, your model shows significantly different performance across different training sets.

- ***How can we solve the issue?***

Q5(b):

During training process, your model shows significantly different performance across different training sets.

- ***How can we solve the issue?***

- Overfitting generally means the model is too complex for the data
    - → Reduce model complexity
    - e.g. via Feature selection, regularisation, tuning hyperparameters
- Can make the data “more complex” → increase training data size

Q6:

Suppose you are given a dataset with single feature *x* and label *y* generated by a function of the form: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

You intend to fit a *regression* model to this data.

If the regression model involves polynomial terms up to x^2, it will likely have:

- Low or high bias?
- Low or high variance?

Q6:

Suppose you are given a dataset with single feature *x* and label *y* generated by a function of the form: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

Model is too simple!

You intend to fit a *regression* model to this data.

If the regression model involves polynomial terms up to x^2, it will likely have:

- Low or high bias?
- Low or high variance?

→ unable to capture true distribution

→ underfit