Exam Revision

Bojan Božić

TU Dublin

April 27, 2020

Introduction

- Date and Time: 11th May 2020, 9am to 6pm.
- Place: Brightspace
- Type: Open Book Exam
- Contents:
  - Basic terms from introduction and evaluation, e.g. supervised learning, bias, model and algorithm (examples and understanding, think practical).
  - Models (solve and explain what you did):
    - Information-based models
    - Similarity-based models
    - Probability-based models
    - Error-based models
  - **NO PLAGIARISM!**

Formulæ

- Shannon's Entropy
- Information Gain
- Euclidean Distance
- kNN
- Normalisation
- Naïve Bayes
- Linear Regression
- Misclassification and Classification Accuracy
- Precision, Recall, F1 Score

Shannon's Entropy

- Shannon's model of entropy is a weighted sum of the logs of the probabilities of each of the possible outcomes when we make a random selection from a set.

$$H(t) = -\sum_{i=1}^{l} (P(t=i) \times log_s(P(t=i))) \tag{1}$$

- What is the entropy of a set of 52 different playing cards?

$$\begin{aligned} H(card) &= -\sum_{i=1}^{52} P(card=i) \times log_2(P(card=i)) \\ &= -\sum_{i=1}^{52} 0.019 \times log_2(0.019) = -\sum_{i=1}^{52} -0.1096 \\ &= 5.700\ bits \end{aligned}$$

Information Gain

Computing information gain involves the following 3 equations:

$$H(t, \mathcal{D}) = -\sum_{l \in levels(t)} (P(t = l) \times log_2(P(t = l))) \tag{2}$$

$$rem(d, \mathcal{D}) = \sum_{l \in levels(d)} \underbrace{\frac{|\mathcal{D}_{d=l}|}{|\mathcal{D}|}}_{\text{weighting}} \times \underbrace{H(t, \mathcal{D}_{d=l})}_{\substack{\text{entropy of} \\ \text{partition } \mathcal{D}_{d=l}}} \tag{3}$$

$$IG(d, \mathcal{D}) = H(t, \mathcal{D}) - rem(d, \mathcal{D}) \tag{4}$$

Euclidean Distance

- One of the best known metrics is Euclidean distance which computes the length of the straight line between two points. Euclidean distance between two instances **a** and **b** in a *m*-dimensional feature space is defined as:

$$Euclidean(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{m} (\mathbf{a}[i] - \mathbf{b}[i])^2} \tag{1}$$

Table: The speed and agility ratings for 20 college athletes labelled with the decisions for whether they were drafted or not.

ID	Speed	Agility	Draft	ID	Speed	Agility	Draft
1	2.50	6.00	No	11	2.00	2.00	No
2	3.75	8.00	No	12	5.00	2.50	No
3	2.25	5.50	No	13	8.25	8.50	No
4	3.25	8.25	No	14	5.75	8.75	Yes
5	2.75	7.50	No	15	4.75	6.25	Yes
6	4.50	5.00	No	16	5.50	6.75	Yes
7	3.50	5.25	No	17	5.25	9.50	Yes
8	3.00	3.25	No	18	7.00	4.25	Yes
9	4.00	4.00	No	19	7.50	8.00	Yes
10	4.25	3.75	No	20	7.25	5.75	Yes

Example

The Euclidean distance between instances d_{12} (SPEED= 5.00, AGILITY= 2.5) and d_5 (SPEED= 2.75,AGILITY= 7.5) in Table 2 [25] is:

$$\begin{aligned} Euclidean(\langle 5.00, 2.50\rangle, \langle 2.75, 7.50\rangle) &= \sqrt{(5.00 - 2.75)^2 + (2.50 - 7.50)^2} \\ &= \sqrt{30.0625} = 5.4829 \end{aligned}$$

k-Nearest Neighbours

- The k nearest neighbors model predicts the target level with the majority vote from the set of k nearest neighbors to the query $\mathbf{q}$:

$$\mathbb{M}_k(\mathbf{q}) = \underset{l \in levels(t)}{\operatorname{argmax}} \sum_{i=1}^{k} \delta(t_i, l) \tag{1}$$

Normalisation

- This odd prediction is caused by features taking different ranges of values, this is equivalent to features having different variances.
- We can adjust for this using normalization; the equation for range normalization is:

$$a_i' = \frac{a_i - min(a)}{max(a) - min(a)} \times (high - low) + low \tag{4}$$

$$x_i.f' = \frac{x_i.f - min(f)}{max(f) - min(f)}$$

Naïve Bayes

Naive Bayes' Classifier

$$\mathbb{M}(\mathbf{q}) = \underset{l \in levels(t)}{\operatorname{argmax}} \left(\prod_{i=1}^{m} P(\mathbf{q}[i] \mid t = l) \right) \times P(t = l)$$

Table: A dataset from a loan application fraud detection domain.

ID	Credit History	Guarantor/ CoApplicant	Accomodation	Fraud
1	current	none	own	true
2	paid	none	own	false
3	paid	none	own	false
4	paid	guarantor	rent	true
5	arrears	none	own	false
6	arrears	none	own	true
7	current	none	own	false
8	arrears	none	own	false
9	current	none	rent	false
10	none	none	own	true
11	current	coapplicant	own	false
12	current	none	own	true
13	current	none	rent	true
14	paid	none	own	false
15	arrears	none	own	false
16	current	none	own	false
17	arrears	coapplicant	rent	false
18	arrears	none	free	false
19	arrears	none	own	false
20	paid	none	own	false

Naïve Bayes

$$
\begin{aligned}
P(fr) &= 0.3 & P(\neg fr) &= 0.7 \\
P(\mathrm{CH} = \textit{'paid'} \mid fr) &= 0.1666 & P(\mathrm{CH} = \textit{'paid'} \mid \neg fr) &= 0.2857 \\
P(\mathrm{GC} = \textit{'none'} \mid fr) &= 0.8334 & P(\mathrm{GC} = \textit{'none'} \mid \neg fr) &= 0.8571 \\
P(\mathrm{ACC} = \textit{'rent'} \mid fr) &= 0.3333 & P(\mathrm{ACC} = \textit{'rent'} \mid \neg fr) &= 0.1429
\end{aligned}
$$

$$\left(\prod_{k=1}^{m} P\left(\mathbf{q}\left[k\right] \mid fr\right)\right) \times P\left(fr\right) = 0.0139$$

$$\left(\prod_{k=1}^{m} P\left(\mathbf{q}\left[k\right] \mid \neg fr\right)\right) \times P(\neg fr) = 0.0245$$

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMODATION	FRAUDULENT
paid	none	rent	*'false'*

Linear Regression

- We can define a multivariate linear regression model as:

$$\begin{aligned}\mathbb{M}_{\mathbf{w}}(\mathbf{d}) &= \mathbf{w}[0] + \mathbf{w}[1] \times \mathbf{d}[1] + \cdots + \mathbf{w}[m] \times \mathbf{d}[m] \quad (7)\\ &= \mathbf{w}[0] + \sum_{j=1}^{m} \mathbf{w}[j] \times \mathbf{d}[j] \quad (8)\end{aligned}$$

$$\begin{aligned}\textsc{Rental Price} = \mathbf{w}[0] &+ \mathbf{w}[1] \times \textsc{Size} + \mathbf{w}[2] \times \textsc{Floor}\\ &+ \mathbf{w}[3] \times \textsc{Broadband Rate}\end{aligned}$$

$$\begin{aligned}\textsc{Rental Price} = -0.1513 &+ 0.6270 \times \textsc{Size}\\ &- 0.1781 \times \textsc{Floor}\\ &+ 0.0714 \times \textsc{Broadband Rate}\end{aligned}$$

$$\begin{aligned}\textsc{Rental Price} &= -0.1513 + 0.6270 \times 690\\ &\quad -0.1781 \times 11 + 0.0714 \times 50\\ &= 434.0896\end{aligned}$$

Misclassification and Classification Accuracy

$$\text{misclassification accuracy} = \frac{(FP + FN)}{(TP + TN + FP + FN)} \qquad (2)$$

$$\text{misclassification accuracy} = \frac{(2 + 3)}{(6 + 9 + 2 + 3)} = 0.25$$

$$\text{classification accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \qquad (3)$$

$$\text{classification accuracy} = \frac{(6 + 9)}{(6 + 9 + 2 + 3)} = 0.75$$

Precision, Recall, F1 Score

$$\text{precision} = \frac{TP}{(TP + FP)} \tag{5}$$

$$\text{recall} = \frac{TP}{(TP + FN)} \tag{6}$$

$$\text{precision} = \frac{6}{(6+2)} = 0.75$$

$$\text{recall} = \frac{6}{(6+3)} = 0.667$$

$$\text{F}_1\text{-measure} = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})} \tag{7}$$

$$\text{F}_1\text{-measure} = 2 \times \frac{\left(\frac{6}{(6+2)} \times \frac{6}{(6+3)}\right)}{\left(\frac{6}{(6+2)} + \frac{6}{(6+3)}\right)} = 0.706$$