Predictive Modelling Tutorial 7 & 8: Generative Models

Dr Liew How Hui

Jan 2021

Tut 6: Generative Models

$$
\begin{aligned}
h_D(\mathrm{x}) &= \operatorname*{argmax}_{j\in\{1,\dots,K\}} \mathbb{P}(Y=j|\mathrm{X}=\mathrm{x}) \\
&= \operatorname*{argmax}_{j\in\{1,\dots,K\}} \frac{\mathbb{P}(\mathrm{X}=\mathrm{x}|Y=j)\mathbb{P}(Y=j)}{\mathbb{P}(\mathrm{X}=\mathrm{x})} \\
&= \operatorname*{argmax}_{j\in\{1,\dots,K\}} \mathbb{P}(\mathrm{X}=\mathrm{x}|Y=j)\mathbb{P}(Y=j) \\
&= \operatorname*{argmax}_{j\in\{1,\dots,K\}} [\ln\mathbb{P}(\mathrm{X}=\mathrm{x}|Y=j)+\ln\mathbb{P}(Y=j)]
\end{aligned} \tag{1}
$$

Tut 6: Generative Models (cont)

Naive Bayes:

$$\mathbb{P}(\mathrm{X} = \mathrm{x} | Y = j) = \prod_{i=1}^{p} \mathbb{P}(X_i = x_i | Y = j)$$

LDA & QDA:

$$\begin{aligned}&\mathbb{P}(\mathrm{X} = \mathrm{x} | Y = j)\\ =&\frac{1}{(2\pi)^{p/2}\sqrt{|\mathrm{C}_j|}} \exp\left\{-\frac{1}{2}(\mathrm{x} - \boldsymbol{\mu}_j)^T \mathrm{C}_j^{-1} (\mathrm{x} - \boldsymbol{\mu}_j)\right\}.\end{aligned}$$

Tutorial 6, Q1

Ahmad would like to construct a model to decide if a day is suitable to play tennis. The table in the next slide shows the results whether to play tennis, based on Outlook, Temperature and Wind, collected by Ahmad.

Using Naïve Bayes approach with Laplace smoothing, predict whether a sunny day with strong wind, 27°C, is suitable to play tennis.

Tutorial 6, Q1 (cont)

Day	Outlook	Temperature	Wind	PlayTennis
D1	Sunny	34	Weak	No
D2	Sunny	32	Strong	No
D3	Overcast	28	Weak	Yes
D4	Rain	22	Weak	Yes
D5	Rain	16	Weak	Yes
D6	Rain	8	Strong	No
D7	Overcast	12	Strong	Yes
D8	Sunny	20	Weak	No
D9	Sunny	10	Weak	Yes
D10	Rain	23	Weak	Yes
D11	Sunny	19	Strong	Yes
D12	Overcast	21	Strong	Yes
D13	Overcast	31	Weak	Yes
D14	Rain	25	Strong	No

FA May 2020 Q2

The testing dataset of an insurance claim is given in Table 2.1. The variables "gender", "bmi", "age_bracket" and "previous_claim" are the predictors and the "claim" is the response.

Table 2.1: The testing data of an insurance claim (randomly sampled with repeated entry).

gender	bmi	age_bracket	previous_claim	claim
female	under_weight	18-30	0	no_claim
female	under_weight	18-30	0	no_claim
male	over_weight	31-50	0	no_claim
female	under_weight	50+	1	no_claim
male	normal_weight	18-30	0	no_claim
female	under_weight	18-30	1	no_claim
male	over_weight	18-30	1	no_claim
male	over_weight	50+	1	claim
female	normal_weight	18-30	0	no_claim
female	obese	50+	0	claim

FA May 2020 Q2 cont

The "gender" is binary categorical data, the "bmi" is a four-value categorical data with values under_weight, normal_weight, over_weight and obese, the "age_bracket" is a three-value categorical data with value "18-30", "31-50" and "50+", the "previous_claim" is a binary categorical data with 0 indicating "no previous claim" and 1 indicating "having a previous claim". The "claim" is a binary response with values "no_claim" (negative class, with value 1) and "claim" (positive class, with value 0).

FA May 2020 Q2 (b)

Write down the mathematical formula for the Naive Bayes model with the predictors and response in Table 2.3. Use the Naive Bayes model trained on the training data from Table 2.3 to **predict** the "claim" of the insurance data in Table 2.1 as well as **evaluating** the performance of the model by calculating the confusion matrix, accuracy, sensitivity, specificity, PPV, NPV of the logistic model.

FA May 2020 Q2 (b) cont

Table 2.3: The training dataset of an insurance claim data for Naive Bayes model.

Obs.	gender	bmi	age_bracket	previous_claim	claim
1	female	obese	50+	1	no_claim
2	female	under_weight	31-50	0	no_claim
3	male	under_weight	31-50	1	no_claim
4	female	over_weight	18-30	1	no_claim
5	female	normal_weight	31-50	0	no_claim
6	female	under_weight	31-50	0	no_claim
7	female	obese	18-30	0	no_claim
8	male	under_weight	50+	1	no_claim
9	female	normal_weight	31-50	0	no_claim
10	male	over_weight	31-50	0	no_claim
11	female	normal_weight	50+	0	claim
12	male	over_weight	31-50	1	claim
13	male	under_weight	31-50	1	claim
14	male	over_weight	31-50	1	claim
15	male	obese	50+	0	claim
16	male	under_weight	50+	0	claim
17	female	obese	31-50	1	claim
18	female	under_weight	50+	1	claim
19	female	normal_weight	50+	1	claim
20	female	under_weight	18-30	1	claim

Note: The default cut-off is 0.5. (4 marks)

FA May 2020 Q2 (c)

Can we compare the logistic regression model in part (a) to the Naive Bayes model in part (b)? Can we say that the logistic regression model is better than the Naive Bayes model solely based on the performance metrics in part (a) and part (b)? Justify your answers with appropriate theory. (2 marks)

Reference: Tutorial Slide 3 on Logistic Regression.

Tutorial 6, Q2

Factory XYZ produces very expensive and high quality golf balls that their qualities are measured in term of curvature and diameter. Result of quality control by experts is given in the table below:

Curvature	Diameter	Result
2.95	6.63	Passed
2.53	7.79	Passed
3.57	5.65	Passed
3.16	5.47	Passed
2.58	4.46	Not Passed
2.16	6.22	Not Passed
3.27	3.52	Not Passed

Tutorial 6, Q2 (cont)

As a consultant to the factory, you get a task to set up the criteria for automatic quality control using LDA model. Then, the manager of the factory also wants to test your criteria upon a new type of golf ball which have curvature 2.81 and diameter 5.46.
[Ref: `https://people.revoledu.com/kardi/tutorial/LDA/Numerical%20Example.html`]

(a) Plot the data with axes of curvature and diameter. Comment on the plot.

Tutorial 6, Q2 (cont)

(b) Write the data into matrix form by separating into "Passed" and "Not Passed".
(c) Compute the prior probability for both classes.
(d) Compute the mean vectors for both classes.
(e) Compute the group covariance matrix.

Tutorial 6, Q2 (cont)

(f) Write down the discriminant functions for both classes.
(g) Transform all the given data into discriminant functions.
(h) Locate the new golf ball in the plot as well as the functions to classify it.
(i) Plot the discriminant line into plot of $\delta_1(X)$ versus $\delta_2(X)$.