THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

MSBD 5012: Machine Learning

Homework 2 Solutions

Assigned: 06/10/18 **Due Date: 20/10/18**

To submit your work, hand it to the instructor on the due date.

Question 1 Consider the following dataset:

Instance	y	x_1	x_2
1	1	0	0
2	1	0	0
3	1	0	1
4	1	0	1
5	0	1	0
6	0	1	0
7	1	1	1
8	0	1	1

(a) Give the Naïve Bayes model for the data. There is no need to use Laplace smoothing, and there is no need to show the process of calculation.

(b) Calculate the posterior probabilities of the Instances 1 and 7 belonging to the two classes according to the model of the previous sub-question. Show the process of calculation.

Solution: (a) $p(y=0)=3/8$, $p(x_1=0|y=0)=0$, $p(x_2=0|y=0)=2/3$, $p(x_1=0|y=1)=4/5$, $p(x_2=0|y=1)=2/5$

(b)

$$\begin{aligned}
p(\mathbf{x}_1|y=0) &= p(x_1=0|y=0)p(x_2=0|y=0)=0 \\
p(\mathbf{x}_1|y=1) &= p(x_1=0|y=1)p(x_2=0|y=1)=8/25 \\
p(y=0|\mathbf{x}_1) &= \frac{p(y=0)p(\mathbf{x}_1|y=0)}{p(y=0)p(\mathbf{x}_1|y=0)+p(y=1)p(\mathbf{x}_1|y=1)}=0 \\
p(y=1|\mathbf{x}_1) &= 1-p(y=0|\mathbf{x}_1)=1.
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{x}_7|y=0) &= p(x_1=1|y=0)p(x_2=1|y=0)=1/3 \\
p(\mathbf{x}_7|y=1) &= p(x_1=1|y=1)p(x_2=1|y=1)=3/25 \\
p(y=0|\mathbf{x}_7) &= \frac{p(y=0)p(\mathbf{x}_7|y=0)}{p(y=0)p(\mathbf{x}_7|y=0)+p(y=1)p(\mathbf{x}_7|y=1)}=\frac{\frac{3}{8}\frac{1}{3}}{\frac{3}{8}\frac{1}{3}+\frac{5}{8}\frac{3}{25}}=\frac{5}{8} \\
p(y=1|\mathbf{x}_7) &= 1-p(y=0|\mathbf{x}_7)=\frac{3}{8}.
\end{aligned}$$

Question 2 Explain how Gaussian discriminant analysis (GDA) is related to logistic regression. Discuss their pros and cons in terms of the distribution characteristics of data and the amount of data.

Solution: GDA reduces to logistic/softmax regression when different classes share the same covariance matrix.

GDA makes stronger assumptions about data distribution than logisti/softmax regression. For parameter estimation, logistic/softamx regression maximizes the conditional log-likelihood, while GDA maximizes the joint log-likelihood.

When the Gaussian assumptions made by GDA are correct, the GDA will need less training data than logistic regression to achieve a certain level of performance. In contrast, by making significantly weaker assumptions, logistic regression is more robust and less sensitive to incorrect modeling assumptions.

Question 3 Discuss other major pros and cons of generative classifiers and discriminative classifiers that are not covered in Question 2.

Solution: Parameter estimation is easier in generative classifiers than in discriminative classifiers. For example, GDA has closed-form formulae for MLE, while logistic regression requires gradient descent to compute MLE.

It is easier to deal with missing data with generative classifiers: Use the EM algorithm during training and marginalization during testing. No principled way to handle missing data with discriminative classifiers.

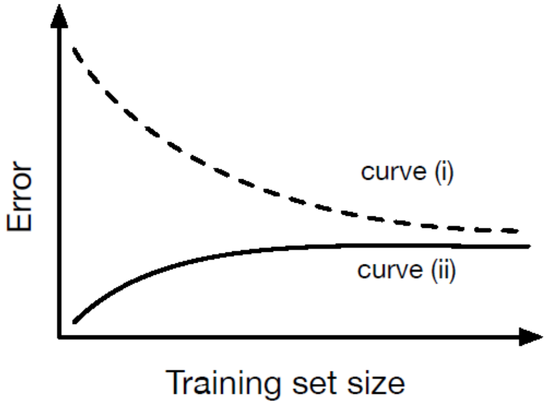
In generative classifiers, we can use unlabeled data to help with the training. There is a subfield called semi-supervised learning that does this. It is much harder to do in discriminative classifiers. In generative classifier, we can do basis function expansion. This cannot be done with generative models.

Feature transformation is usually not used with generative classifiers. One reason is that the transformed features are relative large in number and are strongly correlated. As such, independence assumption is not appropriate. If we do not make independence assumption, we wwill need a large number of parameters, which is quadratic in the number of transformed features. This can easily lead to overfitting. In addition, if the original features are part of the transformed features, the additional features are not useful for classification as they are independent of the class label given the original features.

The aforementioned problems do not apply to discriminative classifiers because they do not model relationships between features. They model only the relationship between features and the class label.

Question 4 The following figure shows the general trend in the training and test errors of classifiers as a function of sample size.

- Which curve represents the training error? Briefly explain.
- What does the gap between the two curves represent? According to the VC Theorem, how does the gap depend on the sample size and model complexity respectively?



Solution: It is easier to correctly classify small training datasets. For example, when the data contains just a single point, it is easy for a classifier have zero training error. On the other hand, we don't expect a classifier learned from few examples to generalize well, so for small training sets the true error is large. Therefore, curve (ii) shows the general trend of the training error.

The gap between the two curves represents the generalization gap. The first part of the VC Theorem states that

With probability at least $1 - \delta$, we have that all for $h \in H$,

$$|\epsilon(h) - \hat{\epsilon}(h)| \leq O(\sqrt{\frac{d}{m} \log \frac{m}{d} + \frac{1}{m} \log \frac{1}{\delta}})$$

This means that the gap $|\epsilon(h) - \hat{\epsilon}(h)|$ increases with model complexity d and decreases with sample size m.