COMP 9501: Machine Learning Spring 2019

Assignment 2

:

1 Problem 1

1.1 Conditional independence and Bayes Ball algorithm

We have discussed in class how to model conditional independence using graphical models, and how to check the conditional Independence between two variable nodes in a graphical model using the Bayes ball algorithm. Answer the questions below, and use either Bayes Ball algorithm or conditional probability to explain. If you use Bayes ball, please describe the path that the 'ball' takes and give an intuition using the canonical three-node graph structures about why (or not) the reachability argument holds, and thus the conditional independence is not satisfied or vice versa.

Considering the model in Figure 1, where the random variables that are conditioning on are not shaded because they change based on the problem.

(a) Is $X_1 \perp X_2 \mid X_3$?

(b) Is $X_1 \perp X_2 \mid X_4$?

(c) Is $X_1 \perp X_2$?

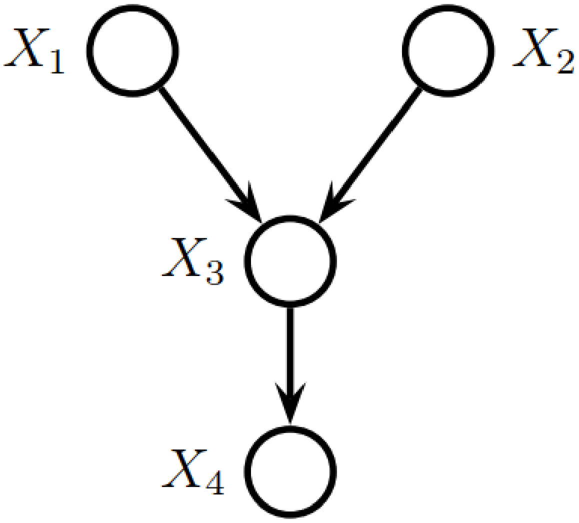

Figure 1: The first graphical model

Considering the model in Figure 2, where the random variables that are conditioning on are not shaded because they change based on the problem.

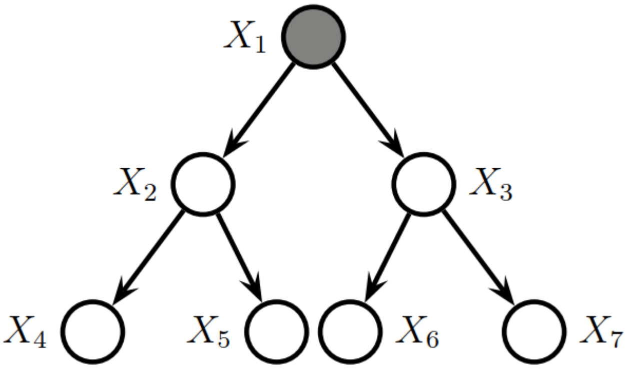

Figure 2: The second graphical model

(d) Is $X_4 \perp X_7 \mid X_1$?

(e) Is $X_4 \perp X_5 \mid X_1$?

1.2 Naive Bayes classifier (NBC)

Naive Bayes classifier is widely used to classify emails. It can be expressed as a graphical model as shown in Figure 3, where Y is the class label, $Y \in \{0, 1\}$. The class 1 means spam email and the class 0 means non-spam email. Random variables $X.$ represent the features of the emails. For instance, X_A can be the number of words in an email, X_B can be the time of day receiving the emails, and X_C can be the number of words not found in dictionary. Let us assume that these three features are Gaussian distributed, although this assumption is not entirely accurate. Let us also assume that Y is Bernoulli with $p(Y = 1|\pi) = \pi$.

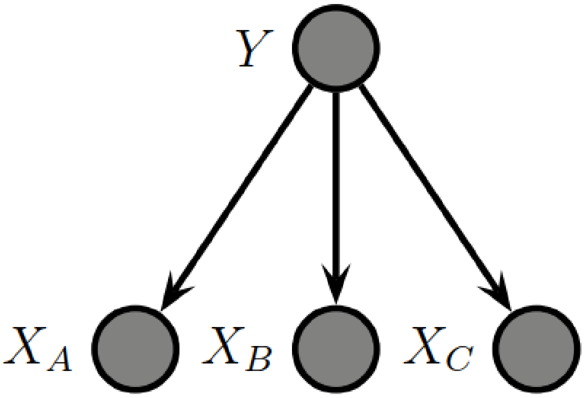

Figure 3: The second graphical model

(a) Write down the joint probability of the random variables in the graph and factorize it according to the graph structure.

(b) For all pairs of features (X_i and X_j, $i \neq j$), is $X_i \perp X_j \mid Y$? Explain what this means with respect to the problem of classification and the implications for the features. Is this true of our real data in practice? Can you think of a situation when this assumption will be explicitly violated, and may impact our classification accuracy? (Two sentences)

(c) We have a training set

$$D_{\text{training}} = \{(Y, X_A, X_B, X_C)_1, \cdots, (Y, X_A, X_B, X_C)_n\}$$

(with observed features and class labels, in other words, n emails classified as spam or not spam), and a test set

$$D_{\text{test}} = \{(X_A, X_B, X_C)_{n+1}, \cdots, (X_A, X_B, X_C)_{n+m}\}$$

(with observed features, but unknown class labels).

With this classifier, we could predict the probability that a test email belongs to the spam class. Write down how to compute $P(Y_j = 1|(X_A, X_B, X_C)_j), j = n+1, \cdots, n+m$, given that we know the components of our factorized joint probability.

1.3 Maximum likelihood estimates for NBC

Still use the graph in Figure 3 and the task of classifying emails. Let's focus on the training data,

$$D_{\text{training}} = \{(Y, X_A, X_B, X_C)_1, \cdots, (Y, X_A, X_B, X_C)_n\}.$$

Suppose our features $X_{A,B,C}$ in each class of emails have Gaussian distributions, e.g. $X_A|Y = y \overset{iid}{\sim} \mathcal{N}(\mu_{A,y}, \sigma^2_{A,y})$. In other words, the Gaussian parameters for the features are class-specific, meaning that there is one mean for feature A when the email is spam and another mean for feature A when it is not spam.

(a) $Y \overset{iid}{\sim} Ber(\pi), \pi \in [0, 1]$. How do we find the MLE of π using the training set D?

(b) If $\sigma^2_{A,y}$ is fixed, derive the MLE of $\mu_{A,1}$ and $\mu_{A,0}$ (in other words, the class mean of feature A for spam and not spam) using training set D.

(c) If $\mu_{A,1}$ and $\mu_{A,0}$ are fixed, derive the MLE for $\sigma^2_{A,0}$ using training set D.

2 Problem 2

We have studied in class how to construct generalized linear models:

$$\xi = \theta^T x, \mu = f(\xi), \eta = \psi(\mu) \tag{1}$$

2.1 Poisson regression

Under this framework, we want to model the data $\{(x_i, y_i)\}$ using Poisson as the conditional distribution of response variables $p(y|x, \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$.

(a) Find the canonical response function $f(\cdot)$,

(b) Derive the batch gradient descent update rule.

(c) Derive the batch Newton descent update rule.

(d) Derive the stochastic gradient descent update rule.

2.2 Multi-class logistic regression

As we discussed in class, we can also use category distribution GLIM to generalize logistic regression to multi-class classification.

(a) Follow what we discussed in class, write down the MLE using the category distribution GLIM.

(b) Derive the batch gradient descent update rule

3 Problem 3

3.1 Linear regression

Some researchers who are desperately in need of a machine learning expert bring you a dataset with information on $n = 1100$ people. Their study has two explanatory predictors: X_1 = a binary indicator of gender (female = 1 and male = 0), and X_2 = some weight. They want to use this information to help predict blood pressure Y which they believe is linearly related to X_1 and X_2.

Suppose that $\sigma^2 = 1$ and for part (c) $\tau^2 = 1$. Use the first 1000 records for your training set, and the last 100 records for your test set. For this answer, include your code (R, Matlab, python, etc.) in your solution, but please do not use built in functions for linear regression. The dataset is HW2_linear_regression.txt.

(a) Write a program to estimate β using the normal equation. Estimate β from the training set.

(b) Write a program to estimate β using stochastic gradient descent. Estimate β from the training set.

(c) Write a program to estimate β using the ridge regression normal equation. Estimate β from the training set.

For all of the above estimation procedures:

(d) Calculate $RSS(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}x_i)^2$ in the training dataset.

(e) Calculate $RSS(\hat{\beta})$ in the test dataset.

(f) For each of the estimated values of $\hat{\beta}$, what is $E[Y|X = [1, 135]^T, \hat{\beta}]$?

3.2 Logistic regression

The researchers this time are interested in doing prediction for a binary outcome Z (an indicator of adverse reaction to a drug they are testing), which they again believe is linearly related to X_1 and X_2.

Again, use the first 1000 records for your training set, and the last 100 records for your test set. The dataset is in HW2_logistic_regression.txt.

(a) Write a program to estimate β using the algorithm introduced in class. Estimate β using the training data.

(b) Calculate $RSS(\hat{\beta})$ in the training dataset.

(c) Calculate $RSS(\hat{\beta})$ in the test dataset.

(d) What is $E[Z|X = [1, 135]^T, \hat{\beta}]$?