Department of Electrical and Computer Engineering

Instituto Superior Técnico

Statistical Learning

2016-2017

Homework 1

All exercises are from: K. Muprhy, "Machine Learning: A Probabilistic Perspective," MIT Press, 2012.

Exercise 2.2 Legal reasoning

(Source: Peter Lee.) Suppose a crime has been committed. Blood is found at the scene for which there is no innocent explanation. It is of a type which is present in 1% of the population.

a. The prosecutor claims: "There is a 1% chance that the defendant would have the crime blood type if he were innocent. Thus there is a 99% chance that he guilty". This is known as the **prosecutor's fallacy**. What is wrong with this argument?
b. The defender claims: "The crime occurred in a city of 800,000 people. The blood type would be found in approximately 8000 people. The evidence has provided a probability of just 1 in 8000 that the defendant is guilty, and thus has no relevance." This is known as the **defender's fallacy**. What is wrong with this argument?

Exercise 2.7 Pairwise independence does not imply mutual independence

We say that two random variables are pairwise independent if

$$p(X_2|X_1) = p(X_2) \tag{2.125}$$

and hence

$$p(X_2, X_1) = p(X_1)p(X_2|X_1) = p(X_1)p(X_2) \tag{2.126}$$

We say that n random variables are mutually independent if

$$p(X_i|X_S) = p(X_i) \quad \forall S \subseteq \{1, \ldots, n\} \setminus \{i\} \tag{2.127}$$

and hence

$$p(X_{1:n}) = \prod_{i=1}^{n} p(X_i) \tag{2.128}$$

Show that pairwise independence between all pairs of variables does not necessarily imply mutual independence. It suffices to give a counter example.

Exercise 4.2 Uncorrelated and Gaussian does not imply independent unless *jointly* Gaussian

Let $X \sim \mathcal{N}(0,1)$ and $Y = WX$, where $p(W=-1) = p(W=1) = 0.5$. It is clear that X and Y are not independent, since Y is a function of X.

a. Show $Y \sim \mathcal{N}(0,1)$.

b. Show $\text{cov}[X,Y] = 0$. Thus X and Y are uncorrelated but dependent, even though they are Gaussian. Hint: use the definition of covariance

$$\text{cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{4.263}$$

and the **rule of iterated expectation**

$$\mathbb{E}[XY] = \mathbb{E}[\mathbb{E}[XY|W]] \tag{4.264}$$

Exercise 5.3 Reject option in classifiers

(Source: (Duda et al. 2001, Q2.13).)

In many classification problems one has the option either of assigning $\mathbf{x}$ to class j or, if you are too uncertain, of choosing the **reject option**. If the cost for rejects is less than the cost of falsely classifying the object, it may be the optimal action. Let α_i mean you choose action i, for $i = 1 : C+1$, where C is the number of classes and $C+1$ is the reject action. Let $Y = j$ be the true (but unknown) **state of nature**. Define the loss function as follows

$$\lambda(\alpha_i|Y=j) = \begin{cases} 0 & \text{if } i = j \text{ and } i,j \in \{1,\ldots,C\} \\ \lambda_r & \text{if } i = C+1 \\ \lambda_s & \text{otherwise} \end{cases} \tag{5.122}$$

In otherwords, you incur 0 loss if you correctly classify, you incur λ_r loss (cost) if you choose the reject option, and you incur λ_s loss (cost) if you make a substitution error (misclassification).

Decision	true label y	
$\hat{y}$	0	1
predict 0	0	10
predict 1	10	0
reject	3	3

a. Show that the minimum risk is obtained if we decide $Y = j$ if $p(Y=j|\mathbf{x}) \geq p(Y=k|\mathbf{x})$ for all k (i.e., j is the most probable class) *and* if $p(Y=j|\mathbf{x}) \geq 1 - \frac{\lambda_r}{\lambda_s}$; otherwise we decide to reject.

b. Describe qualitatively what happens as λ_r/λ_s is increased from 0 to 1 (i.e., the relative cost of rejection increases).

In the following exercise, make your life simpler by assuming that $\bar{y} = 0$ and $\bar{\boldsymbol{x}} = \mathbf{0}$.

Exercise 7.9 Generative model for linear regression

Linear regression is the problem of estimating $E[Y|\mathbf{x}]$ using a linear function of the form $w_0 + \mathbf{w}^T\mathbf{x}$. Typically we assume that the conditional distribution of Y given $\mathbf{X}$ is Gaussian. We can either estimate this conditional Gaussian directly (a discriminative approach), or we can fit a Gaussian to the joint distribution of $\mathbf{X}, Y$ and then derive $E[Y|\mathbf{X} = \mathbf{x}]$.

In Exercise 7.5 we showed that the discriminative approach leads to these equations

$$E[Y|\mathbf{x}] = w_0 + \mathbf{w}^T\mathbf{x} \tag{7.109}$$

$$w_0 = \overline{y} - \overline{\mathbf{x}}^T\mathbf{w} \tag{7.110}$$

$$\mathbf{w} = (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y}_c \tag{7.111}$$

where $\mathbf{X}_c = \mathbf{X} - \bar{\mathbf{X}}$ is the centered input matrix, and $\bar{\mathbf{X}} = \mathbf{1}_n\overline{\mathbf{x}}^T$ replicates $\overline{\mathbf{x}}$ across the rows. Similarly, $\mathbf{y}_c = \mathbf{y} - \overline{\mathbf{y}}$ is the centered output vector, and $\overline{\mathbf{y}} = \mathbf{1}_n\overline{y}$ replicates $\overline{y}$ across the rows.

a. By finding the maximum likelihood estimates of $\boldsymbol{\Sigma}_{XX}$, $\boldsymbol{\Sigma}_{XY}$, $\boldsymbol{\mu}_X$ and $\boldsymbol{\mu}_Y$, derive the above equations by fitting a joint Gaussian to $\mathbf{X}, Y$ and using the formula for conditioning a Gaussian (see Section 4.3.1). Show your work.

b. What are the advantages and disadvantages of this approach compared to the standard discriminative approach?

The datasets and code mentioned in the next exercise are available at

https://code.google.com/p/pmtkdata/ and https://github.com/probml/pmtk3

Exercise 8.1 Spam classification using logistic regression

Consider the email spam data set discussed on p300 of (Hastie et al. 2009). This consists of 4601 email messages, from which 57 features have been extracted. These are as follows:

- 48 features, in $[0, 100]$, giving the percentage of words in a given message which match a given word on the list. The list contains words such as "business", "free", "george", etc. (The data was collected by George Forman, so his name occurs quite a lot.)
- 6 features, in $[0, 100]$, giving the percentage of characters in the email that match a given character on the list. The characters are ; ([! $ #
- Feature 55: The average length of an uninterrupted sequence of capital letters (max is 40.3, mean is 4.9)
- Feature 56: The length of the longest uninterrupted sequence of capital letters (max is 45.0, mean is 52.6)
- Feature 57: The sum of the lengts of uninterrupted sequence of capital letters (max is 25.6, mean is 282.2)

Load the data from `spamData.mat`, which contains a training set (of size 3065) and a test set (of size 1536).

One can imagine performing several kinds of preprocessing to this data. Try each of the following separately:

a. Standardize the columns so they all have mean 0 and unit variance.

b. Transform the features using $\log(x_{ij} + 0.1)$.

c. Binarize the features using $\mathbb{I}(x_{ij} > 0)$.

For each version of the data, fit a logistic regression model. Use cross validation to choose the strength of the ℓ_2 regularizer. Report the mean error rate on the training and test sets. You should get numbers similar to this:

method	train	test
stnd	0.082	0.079
log	0.052	0.059
binary	0.065	0.072