Unsupervised Learning and Prediction

Unsupervised learning can play an important role in prediction,

both for regression and classification problems. In some cases, we

want to predict a category in the absence of any labeled data. For

example, we might want to predict the type of vegetation in an area

from a set of satellite sensory data. Since we don’t have a response

variable to train a model, clustering gives us a way to identify common

patterns and categorize the regions.

Clustering is an especially important tool for the “cold-start problem.”

In this type of problem, such as launching a new marketing

campaign or identifying potential new types of fraud or spam, we

initially may not have any response to train a model. Over time, as

data is collected, we can learn more about the system and build a

traditional predictive model. But clustering helps us start the learning

process more quickly by identifying population segments.

Unsupervised learning is also important as a building block for

regression and classification techniques. With big data, if a small

subpopulation is not well represented in the overall population, the

trained model may not perform well for that subpopulation. With

clustering, it is possible to identify and label subpopulations. Separate

models can then be fit to the different subpopulations. Alternatively,

the subpopulation can be represented with its own feature,

forcing the overall model **B**

It is also common to compute principal components on deviations

from the means of the predictor variables, rather than on the values

themselves. **B**

The weights for the first principal component are both negative,

but reversing the sign of all the weights does not change the principal

component. For example, using weights of 0.747 and 0.665 for

the first principal component is equivalent to the negative weights,

just as an infinite line defined by the origin and 1,1 is the same as

one defined by the origin and –1, –1. **B**

How Many Components to Choose?

If your goal is to reduce the dimension of the data, you must decide

how many principal components to select. The most common

approach is to use an ad hoc rule to select the components that

explain “most” of the variance. You can do this visually through the

screeplot, as, for example, in Figure 7-2. Alternatively, you could

select the top components such that the cumulative variance

exceeds a threshold, such as 80%. Also, you can inspect the loadings

to determine if the component has an intuitive interpretation.

Cross-validation provides a more formal method to select the

number of significant components (see “Cross-Validation” on page

155 for more). **B**

Normalization

It is typical to normalize (standardize) continuous variables by subtracting

the mean and dividing by the standard deviation. Otherwise,

variables with large scale will dominate the clustering process

(see “Standardization (Normalization, z-Scores)” on page 243). **B**

Cluster Mean

In clustering records with multiple variables (the typical case), the

term *cluster mean* refers not to a single number but to the vector of

means of the variables. **O**

Cluster Analysis Versus PCA

The plot of cluster means is similar in spirit to looking at the loadings

for principal components analysis (PCA); see “Interpreting

Principal Components” on page 289. A major distinction is that

unlike with PCA, the sign of the cluster means is meaningful. PCA

identifies principal directions of variation, whereas cluster analysis

finds groups of records located near one another. **B**

There are several more formal ways to determine the number of

clusters based on statistical or information theory. For example,

Robert Tibshirani, Guenther Walther, and Trevor Hastie propose a

“gap” statistic based on statistical theory to identify the elbow. For

most applications, a theoretical approach is probably not necessary,

or even appropriate. **B**

Note that in most cases BIC is usually minimized. The authors of

the mclust package decided to define BIC to have the opposite sign

to make interpretation of plots easier.**O**

Model-based clustering is a rich and rapidly developing area of

study, and the coverage in this text spans only a small part of the

field. Indeed, the mclust help file is currently 154 pages long. Navigating

the nuances of model-based clustering is probably more

effort than is needed for most problems encountered by data

scientists. **B**

Scaling is also important for PCA. Using the *z*-scores is equivalent

to using the correlation matrix (see “Correlation” on page 30)

instead of the covariance matrix in computing the principal components.

Software to compute PCA usually has an option to use the

correlation matrix (in *R*, the princomp function has the argument

cor). **B**