email message and characters

_		
е	edu	remove
3	0.01	0.28
2	0.29	0.01

ouild a predicie outcome for redicts such an

learning proboutcome varirning problem, f the outcome. r clustered. We rvised problem st chapter. re discussed in

501 email mesjunk email, or detector that s. For all 4601 am is available, nonly occurring is a supervised /spam. It is also

largest average

se and how: for

1.

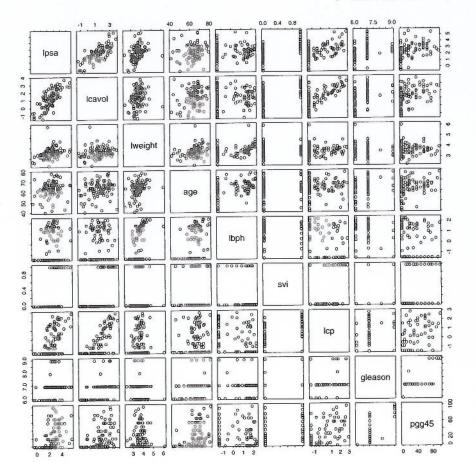


FIGURE 1.1. Scatterplot matrix of the prostate cancer data. The first row shows the response against each of the predictors in turn. Two of the predictors, svi and gleason, are categorical.

For this problem not all errors are equal; we want to avoid filtering out good email, while letting spam get through is not desirable but less serious in its consequences. We discuss a number of different methods for tackling this learning problem in the book.

Example 2: Prostate Cancer

The data for this example, displayed in Figure 1.1, come from a study by Stamey et al. (1989) that examined the correlation between the level of prostate specific antigen (PSA) and a number of clinical measures, in 97 men who were about to receive a radical prostatectomy.

The goal is to predict the log of PSA (lpsa) from a number of measurements including log-cancer-volume (lcavol), log prostate weight lweight,

vasion svi, log of capsular penetration 1cp, Gleason score gleason, and percent of Gleason scores 4 or 5 pgg45. Figure 1.1 is a scatterplot matrix age, log of benign prostatic hyperplasia amount 1bph, seminal vesicle inof the variables. Some correlations with 1psa are evident, but a good predictive model is difficult to construct by eye.

This is a supervised learning problem, known as a regression problem, because the outcome measurement is quantitative.

Example: Prostate Cancer 3.2.1

showing every pairwise plot between the variables. We see that svi is a weight (lweight), age, log of the amount of benign prostatic hyperplasia strong correlations. Figure 1.1 (page 3) of Chapter 1 is a scatterplot matrix a number of clinical measures in men who were about to receive a radical prostatectomy. The variables are log cancer volume (lcavol), log prostate The correlation matrix of the predictors given in Table 3.1 shows many The data for this example come from a study by Stamey et al. (1989). They examined the correlation between the level of prostate-specific antigen and (1bph), seminal vesicle invasion (svi), log of capsular penetration (1cp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45)

ual genes, a red (positiv ample of Fi although fo ure displays experiment.

The chall ganized. Ty

sion I (a) which

profile (b) which

(c) do ceri

samp

We could predictor v the level of

unsupervisa think of th to cluster t