10-702 Statistical Machine Learning: Assignment 4

Due Friday, March 21

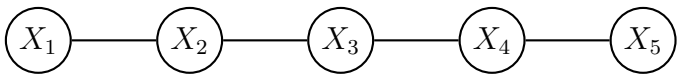
(1) Let $X = (X_1, \ldots, X_d)^T$ where each $X_j \in \{0, 1\}$. Consider the loglinear model

$$\log p(x) = \beta_0 + \sum_{j=1}^{d} \beta_j X_j + \sum_{j<k}^{d} \beta_{jk} X_j X_k + \cdots + \sum_{j<k<\ell}^{d} \beta_{jk\ell} X_j X_k X_\ell + \cdots +$$

Suppose that $\beta_A = 0$ whenever $\{1, 2\} \subset A$. Show that

$$X_1 \amalg X_2 \,\Big|\, X_3, \ldots, X_d.$$

(2) Consider the graph:



Assume all variables are binary. One loglinear model that is consistent with this graph is

$$\log p(x) = \beta_0 + 5\Big(X_1X_2 + X_2X_3 + X_3X_4 + X_4X_5\Big).$$

Simulate $n = 100$ random vectors from this distribution. Fit the model

$$\log p(x) = \beta_0 + \sum_{j} \beta_j X_j + \sum_{j<k} \beta_{jk} X_j X_k$$

using maximum likelihood. Report your estimators. Use forward model selection with BIC to choose a submodel. Compare the selected model to the true model.

(3) Let $X \sim N(0, \Sigma)$ where $X = (X_1, \ldots, X_p)^T$, $p = 10$. Let $\Theta = \Sigma^{-1}$ and suppose that $\Theta(i,i) = 1$, $\Theta(i, i-1) = .5$, $\Theta(i-1, i) = .5$ and $\Theta(i,j) = 0$ otherwise. Simulate 50 random vectors and use the glasso to estimate the covariance matrix. Compare your estimated graph to the true graph.

(4) Generate $n = 100$ observations from the model:

$$Y = m(X) + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$, $\sigma = 0.1$,

$$X \sim \text{Uniform}([0,1]^{10})$$

and

$$m(x) = \cos(5\pi x_1) + 5x_2^2.$$

Note that x is 10-dimensional but $m(x)$ only depends on x_1 and x_2. Estimate m using (i) multivariate kernel regression, (ii) additive model, (iii) regression tree. For each estimator, report

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{m}(X_i) - m(X_i))^2.$$

(5) You're goal is to compare several classifiers, namely: (i) logistic regression, (ii) additive model, (iii) k-nearest neighbors and (iv) classification trees. The data are the "iris data" which is a famous dataset. These data are already in R. Type:

```
data(iris)
names(iris)
print(iris)
pairs(iris)
boxplot(iris)
```

There are three species and the goal is to predict species from the four features. We will use the first 100 observations only so there are only two species. Construct the classifiers and compare the training error rates. For the additive model, plot the estimated functions.