SYS 6003 - Homework 9

Due: Thursday, 11/29/2012 in class

1. *Predicting football outcomes*

   In this problem you will build a simple predictive model for the outcomes of football games. This model attempts to quantify the following intuition:

   - The strength of a team's offense can be measured by a statistic related to the number of points this team scores per game. That is, if a team tends to score many points per game, this team is likely to win future games by outscoring their opponents.
   - The strength of a team's defense can be measured by a statistic related to the number of points scored against this team per game. That is, if a team tends to prevent their opponents from scoring many points, this team is likely to win future games by keeping their opponents' scores lower than their own.
   - One simple way to measure offensive and defensive strength would be to use the average points scored and average points allowed per game. However, these statistics disregard the strength of the teams played in previous games. For example, scoring an average of 21 points per game against teams with strong defenses might be a better indicator of offensive strength than scoring an average of 24 points per game against teams with weak defenses.

   We will suggest a simple probabilistic model that will be used to explain scoring patterns, then use maximum likelihood estimation to estimate the parameters of this model. Let Z_{ij} be a random variable indicating the number of points that team i scores when playing team j. We will model this random variable as Gaussian[1] with mean and variance given by

   $$\mathbf{E}[Z_{ij}] = \frac{\mu_i}{\alpha_j} \qquad \text{and} \qquad \mathbf{Var}[Z_{ij}] = \frac{\sigma^2}{\alpha_j^2}$$

   Here μ_i is a parameter that is large if team i is a high scoring team, α_j is a parameter that is large if team j prevents their opponents from scoring, and σ is a fixed parameter that we will use for all all games. So, the probability density function of Z_{ij} is given by

   $$f_{ij}(z \,|\, \mu_i, \alpha_j) = \frac{\alpha_j}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(z - \mu_i/\alpha_i)^2}{(\sigma^2/\alpha_j^2)}\right)$$

[1] Of course this is a very crude model. Football scores are integer valued and nonnegative. However, this model is analytically convenient and will still allow us to detect useful patterns in our historical data.

The goals of this problem are to use historical football scores to estimate values of μ_i and α_i for all 32 teams in the National Football League, then use the resulting models to attempt to predict the winners of all 16 football games that will be played between 11/22/2012–11/26/2012.

To identify each team, each of the 32 teams in the NFL will be assigned an index in the set $\{1, \ldots, 32\}$. As of 11/20/2012, 160 regular-season games have been played in the 2012 NFL season. These games will be indexed by $k \in \{1, \ldots, 160\}$. Let i_k be the index of the home team in game k, and j_k be the index of the visitor in game k. Let $z_{i_k j_k}$ be the number of points that the home team scored against the visitor in game k and $z_{j_k i_k}$ be the number of points that the visitor scored against the home team in game k. Finally, assume that the scoring outcomes of all games are conditionally independent given the μ and α parameters for all teams.

According to our model, the likelihood of observing all scores for the past 160 games is

$$\prod_{k=1}^{160} f_{i_k j_k}(z_{i_k j_k} \,|\, \mu_{i_k}, \alpha_{j_k}) f_{j_k i_k}(z_{j_k i_k} \,|\, \mu_{j_k}, \alpha_{i_k})$$

(a) Show that the problem of selecting the μ_i and α_i for all teams that maximizes the likelihood above is equivalent to selecting the μ_i and α_i for all teams that minimizes

$$\frac{1}{2\sigma^2} \sum_{k=1}^{160} (\mu_{i_k} - \alpha_{j_k} z_{i_k j_k})^2 + \frac{1}{2\sigma^2} \sum_{k=1}^{160} (\mu_{j_k} - \alpha_{i_k} z_{j_k i_k})^2 - \sum_{i=1}^{32} N_i \ln(\alpha_i),$$

where N_i indicates the number of games played by team i.

(b) The Matlab file `hw9.mat` contains a matrix called `scores`, which contains all scores for the past 160 games (details of all Matlab data for this problem are provided on the next page). Use this data to compute the μ_i and α_i values for each team that minimize the objective function above. Provide your Matlab code and all 32 of your μ_i and α_i values.

(c) In a game between team i and team j, our model predicts that team i is more likely to win if

$$\frac{\mu_i}{\sigma_j} > \frac{\mu_j}{\sigma_i}$$

Use the model to predict winners for all 16 games given in the matrix `games`. Provide your Matlab code and all 16 of your predicted winners.

The Matlab file `hw9.mat` contains the following data:

- The matrix `teams` contains the names of the 32 NFL teams. The index of a team is given by the team's position in this matrix.
- The matrix `scores` contains all scores for the 160 games that have been played in the 2012 NFL season. Each row contains the matchup and final score of a game. Column 1 contains the index of the visiting team. Column 2 contains the index of the home team. Column 3 contains the number of points scored by the visiting team. Column 4 contains the number of points scored by the home team.
- The matrix `games` contains all matchups for the 16 games to be played between 11/22/2012–11/26/2012. Each row contains a pair of teams that is playing a game. Column 1 contains the index of the visiting team. Column 2 contains the index of the home team.