Problem 1 (Classifying forest fire damages)

The forestfires.csv files on the class website contains 517 instances of forest fires in Montesinho Natural Park, Portugal. The forestfiresmeta.txt file contains some metadata. Among other variables, the data include

- **FFMC** = fine fuel moisture code, a numerical rating of the moisture content of surface litter
- **DMC** = Duff moisture code, a numerical rating of the average moisture content of loosely compacted organic layers of moderate depth
- **DC** = drought code, a numerical rating of the moisture content of deep, compact, organic layers.
- **ISI** = initial spread index, based on FFMC and wind, measures the rate a fire will spread in its early stages

while all other variables are described in the metadata.

Your goal in this problem is to carefully select and compare a support vector machine against a logistic regression for predicting whether a fire will cause damage, that is, your goal is to predict the two classes \( \text{area} = 0 \) vs. \( \text{area} > 0 \). Be sure to initially split the data into training and testing, and report relevant predictive summaries for the two models on the testing data.

Turn in your code as an Appendix.

Problem 2 (Support vector machines two ways)

The HW5.RData file on the website contains a set of simulated data with outcomes \( y \in \{-1, +1\} \) and two real-valued features \( X_1 \) and \( X_2 \). For this problem fit two models:

- **Model 1**: A support vector classifier with prediction features \( X_1, X_2, X_1^2, X_2^2, X_1X_2 \). Use \texttt{cost=3.4} in the \texttt{svm} function. (Hint: you will need to create a new data frame with three extra columns for the nonlinear terms to pass into \texttt{svm}).

- **Model 2**: A support vector machine using a polynomial kernel of degree 2 with features \( X_1 \) and \( X_2 \). Use \texttt{cost=0.62} in the \texttt{svm} function.

Both costs were already found by 10-fold cross-validation. Compare the two models by plotting classification regions (say by evaluating predictions on a fine two-dimensional grid) for the two models. Do they differ, or are they the same? Does either decision boundary look linear?
Problem 3 (Grad students only)

Suppose $Y = X\beta + \epsilon$ where $\epsilon$ consists of uncorrelated, mean zero random variables each with variance $\sigma^2$ (e.g., any of assumptions A1, A2 or A3). For a smoothing parameter $\lambda > 0$, let

$$\hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1}X^TY$$

be the ridge regression estimate of $\beta$. In this problem you will derive the mean squared error (MSE) of $\hat{\beta}_{ridge}$, and compare it to the MSE of the ordinary least squares estimator.

(a) Suppose $Z$ is a random vector with mean $\mu$ and covariance matrix $\Sigma$. If $M$ is a matrix, show that

$$E(Z^T MZ) = \mu^T M \mu + \text{tr}(M\Sigma).$$

(b) Show

$$MSE(\lambda) = E((\beta - \hat{\beta}_{ridge})^T (\beta - \hat{\beta}_{ridge}))$$

$$= \beta^T (I - M(\lambda)X)^T (I - M(\lambda)X) \beta + \sigma^2 \text{tr}(M(\lambda)^T M(\lambda))$$

where $M(\lambda) = (X^TX + \lambda I)^{-1}X^T$.

(c) What does this simplify to for the OLS estimator, i.e., $\lambda = 0$?

(d) Simplify parts (b) and (c) for $p = 1$, i.e., only one feature: $Y = \beta_1 X_1 + \epsilon$. Are there any choices of $\lambda$ for which the MSE under ridge regression is strictly less than the MSE under OLS?

Problem 4 (Grad students only)

The paper by Boser, Guyon and Vapnik (1992) first introduced the notion of a support vector machine by using kernels in the decision function. Read it.