**STA 414S/2104S**: TakeHome Midterm Test. Due March 25, 2010 before 2 pm.

**Please work alone.**

1. (Adapted from Exercise 2.7, HTF). Suppose we have a sample $(y_1, x_1), \ldots, (y_N, x_N)$, and we assume the model
$$y_i = f(x_i) + \epsilon_i, \tag{1}$$
where $f(\cdot)$ is an unkown regression function, $\epsilon_i \sim N(0, \sigma^2)$, and the $\epsilon$'s are independent. A fairly wide class of estimators considered in the course are of the form
$$\hat{f}(x_0) = \sum_{i=1}^{N} \ell_i(x_0; \mathbf{x}) y_i,$$
where $\mathbf{x} = (x_1, \ldots, x_N)$.

   (a) Show that linear regression and $k$-nearest neighbour regression are members of this class of estimators, and describe the weights $\ell_i(x_0; \mathbf{x})$ in each of these cases.

   (b) **STA 2104 only** Decompose the conditional mean-squared error
$$E_{\mathbf{y}|\mathbf{x}}\{\hat{f}(x_0) - f(x_0)\}^2,$$
where the expectation is over the conditional distribution of $y_1, \ldots, y_N$, given $x_1, \ldots, x_N$.

2. (Adapted from Exercise 2.1, R.A. Berk). Figure 1 shows a plot of Ozone against Temperature, from a database of daily measurements in New York over 154 summer days. The following fits are summarized in the code extract:

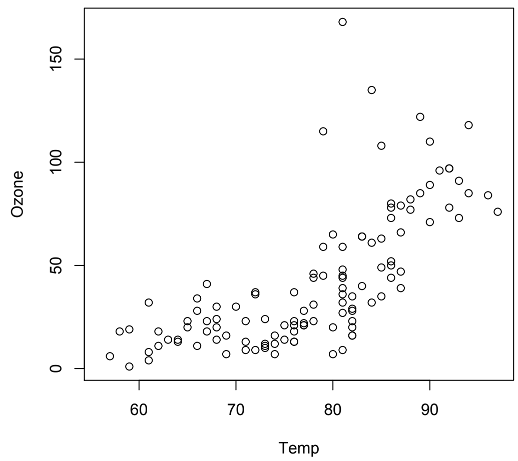
```
> data(airquality);attach(airquality)
> library(gam)
> out1 = gam(Ozone ~ Temp)
> out2 = gam(Ozone ~ as.factor(Temp))
> out3 = gam(Ozone ~ s(Temp))  # the degrees of freedom
                               # for s(.) will be selected by GCV
```

   (a) The first model is the smoothest possible model; the second is the roughest possible model, and the third is somewhere in between.[1] Explain why this is the case.

   (b) A summary of the output is presented below. Which model has the best fit judging by the residual deviance? Which model has the best fit judging by the AIC? Why might the choice of the best model differ depending on which measure of fit is used? Which model seems to be the most useful judging from Figure 2?

[1] The function `gam` assumes Normal errors if nothing is specified, so the model for `out1`, for example, can also be fit using `lm(Ozone ~ Temp)`.

Figure 1: Plot of Ozone vs. Temperature, using dataset `airquality`.




(c) **STA 2104 only** Experiment with `out3` with different amounts of smoothing, and present plots that include a confidence band for the smooth function. Explain how this confidence band was computed.

```
> summary(out1)

Call: gam(formula = Ozone ~ Temp)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-40.7295 -17.4086  -0.5869  11.3062 118.2705

(Dispersion Parameter for gaussian family taken to be 562.3675)

    Null Deviance: 125143.1 on 115 degrees of freedom
Residual Deviance: 64109.89 on 114 degrees of freedom
AIC: 1067.706
37 observations deleted due to missingness

Number of Local Scoring Iterations: 2
DF for Terms


            Df
(Intercept)  1
Temp         1


> summary(out2)

Call: gam(formula = Ozone ~ as.factor(Temp))
Deviance Residuals:
   Min     1Q Median     3Q    Max
```

```
 -44.0   -9.0   -1.0    8.0  117.3

(Dispersion Parameter for gaussian family taken to be 500.02)

    Null Deviance: 125143.1 on 115 degrees of freedom
Residual Deviance: 38501.54 on 77 degrees of freedom
AIC: 1082.558
37 observations deleted due to missingness

Number of Local Scoring Iterations: 2
DF for Terms


                Df
(Intercept)      1
as.factor(Temp) 38
> summary(out3)

Call: gam(formula = Ozone ~ s(Temp))
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-34.924 -11.144  -3.414   8.595 124.076

(Dispersion Parameter for gaussian family taken to be 483.8819)

    Null Deviance: 125143.1 on 115 degrees of freedom
Residual Deviance: 53710.96 on 111.0001 degrees of freedom
AIC: 1053.176
37 observations deleted due to missingness

Number of Local Scoring Iterations: 2

DF for Terms and F-values for Nonparametric Effects

            Df Npar Df Npar F     Pr(F)
(Intercept)  1
s(Temp)      1       3 7.1639 0.0001929 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

3. This question requires you to read the paper by M. Zhu, "Kernels and ensembles: perspectives on statistical learning". It was handed out in class and is also posted on the web page. (You may omit §4.)

   (a) Explain in two or three sentences, without using equations, the "kernel trick", as applied to support vector machines.

   (b) Explain in two or three sentences, without using equations, the method of ensembles. Zhu describes this method as "relatively mindless": why?

   (c) Compare mis-classification rates for the classification problem for the wine data of HW 2 using support vector machines and random forests, and compare these to each other and to your results from HW 2.

Figure 2: Data with fitted values from `out1` , `out2` , `out3`