Math 650 HW 2

Mose Wintner

1. First, $1 - p(X) = \frac{1}{1+e^{\beta_0+\beta_1 X}}$. So after clearing denominators,

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}.$$

2. Starting with (4.12), we cancel the $\frac{1}{\sigma\sqrt{2\pi}}$. Since log is an increasing function, maximizing $p_k(x)$ over k is equivalent to maximizing its logarithm,

$$\log(\pi_k) - \frac{1}{2\sigma^2}(x^2 - 2x\mu_k + \mu_k^2) - \log\left(\sum_{\ell=1}^{K} \pi_\ell \exp\left(-\frac{1}{2\sigma^2}(x-\mu_\ell)^2\right)\right)$$

where we've expanded the quadratic middle term. Maximizing $p_k(x)$ over k doesn't depend on the last term, nor the x^2 term, so maximizing $p_k(x)$ over k given x is equivalent to maximizing

$$\log(\pi_k) - \frac{1}{2\sigma^2}(-2x\mu_k + \mu_k^2) = \log(\pi_k) - \frac{\mu_k^2}{2\sigma^2} + x\frac{\mu_k}{\sigma^2}$$

3. The likelihood of being drawn from distribution k is

$$p_k(x) = \frac{\pi_k \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}{\sum_{\ell=1}^{K} \pi_\ell \frac{1}{\sigma_\ell\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_\ell^2}(x-\mu_\ell)^2\right)}$$

We can cancel the $\frac{1}{\sqrt{2\pi}}$. Maximizing this quantity over k is again equivalent to maximizing its logarithm. Again, the denominator (and its logarithm) do not depend on k so we can ignore it.

$$\log(\pi_k) - \log(\sigma_k) - \frac{1}{2\sigma_k^2}(x-\mu_k)^2 = -\frac{1}{2\sigma_k^2}x^2 + \frac{\mu_k}{\sigma_k^2}x - \frac{\mu_k^2}{\sigma_k^2} + \log(\pi_k) - \log(\sigma_k)$$

which is quadratic in x.

4. a) 0.1
b) $(0.1)^2$
c) $(0.1)^{100}$
d) Suppose we fix a notion of "near" in one dimension, and decide that the notion of "near" in higher dimensions is "near in all projections onto the features' (standardized) coordinate axes." Then the number of training observations "near" a test observation grows exponentially with p.
e) For $p = 1$, the iide length is 0.1. For $p = 2$, it's $\sqrt{(0.1)} \approx 0.31$. For $p = 100$, it's $(0.1)^{1/100} \approx 0.977$. This is not useful, because each "side" of the cube corresponds to a single feature, so an observation's falling outside this hypercube could be the result of one "unlucky" observation of a single feature.

5. a) We'd expect LDA to generally perform OK on training data and pretty well on test data, because LDA estimates the Bayes decision boundaries as linear functions of x. Therefore if the

Bayes decision boundary is linear, LDA should outperform QDA. It won't perform spectacularly on training data when compared with QDA, but to use QDA would be to overfit, since it's too flexible a method.
b) We'd expect QDA to peform decently well on training data and on test data, because a nonlinear curve is better approximated by a parabola than a line.
c) We'd expect QDA to perform better in general, because as a more flexible method, it can better fit a large amount of data. Also, if the decision boundary is truly linear, QDA will still end up doing an OK job of approximating it by training a small quadratic coefficient; on the other hand, LDA can't approximate a nonlinear boundary well.
d) FALSE. The variance of QDA will still be higher than that of LDA because the boundary is actually linear, and in practice the QDA boundary will only be *nearly* linear at best. Using LDA to approximate a linear boundary will avoid error due to curvature.

6. a)

$$\frac{e^{-6+.05\cdot 40+3.5}}{1+e^{-6+.05\cdot 40+3.5}} \approx 0.378.$$

b)

$$\begin{aligned}
.5 &= \frac{e^{-6+.05h+3.5}}{1+e^{-6+0.5h+3.5}} && (1)\\
.5(1+e^{.05h-2.5}) &= e^{.05h-2.5} && (2)\\
1 &= e^{.05h-2.5} && (3)\\
2.5 &= .05h && (4)\\
h &= 50 && (5)
\end{aligned}$$

so, 50 hours.

7.

$$\begin{aligned}
P(\text{dividend this year} \mid X = 4 \text{ last year}) &= \frac{P(\text{dividend this year}, X = 4 \text{ last year})}{P(X = 4 \text{ last year})}\\
&= \frac{\frac{0.8}{\sqrt{12\pi}}e^{-(4-10)^2/(2\cdot 36)}}{\frac{0.8}{\sqrt{12\pi}}e^{-(4-10)^2/(2\cdot 36)} + \frac{0.2}{\sqrt{12\pi}}e^{-4^2/(2\cdot 36)}}\\
&= \frac{4e^{-1/2}}{4e^{-1/2}+e^{-2/9}}\\
&\approx 0.75
\end{aligned}$$

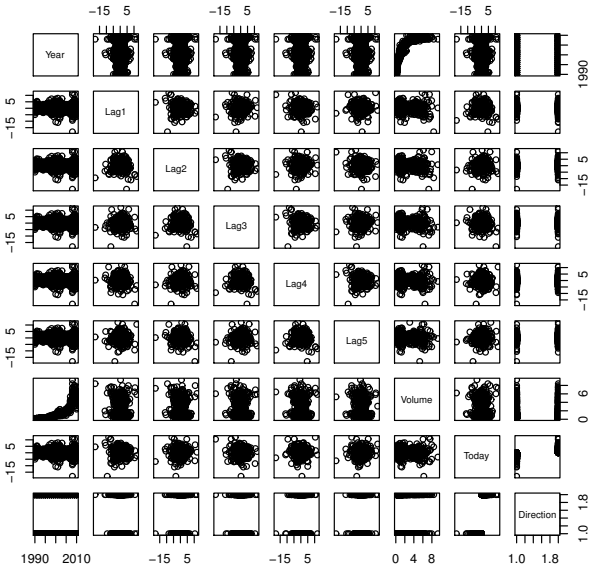
8. 1NN is prone to overfitting, and logistic regression is fairly rigid. Therefore, in absence of more information about the problem, performance on training data should be ignored for purposes of determining the best method. Suppose there are 100 data points, so 50 training and 50 test. Then 15 test values were misclassified under logistic regression. However, all 18 1NN errors were test errors, because there is no training error in 1NN. Each training point is its own nearest neighbor. Therefore the logistic regression performed better on the test set, so we should probably choose it over 1NN.

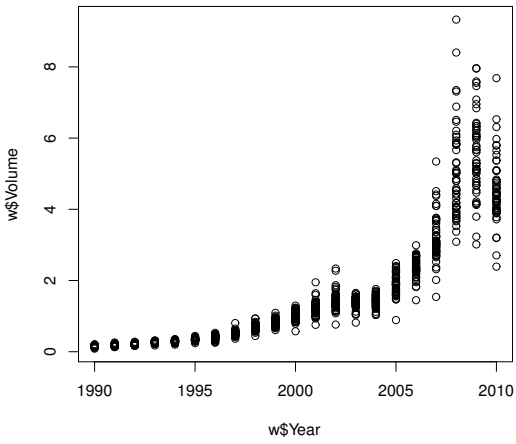
9. a)

$$\frac{p}{1-p} = 0.37$$
$$p = .37 - .37p$$
$$1.37p = .37$$
$$p = 0.2701$$

b) It's $.16/.84 = 0.1905$.

10. a) The quantities Today and Lag1 through Lag5 all appear to have very close to the same distribution, from summary data. The plots of a day's market vs. another day's market appear to mimic a Gaussian distribution around $(0, 0)$. The only variables that appear to be correlated are Volume and Year; Volume appears to be increasing exponentially over time, which is as expected. The correlation matrix appears to be relatively close to the identity, meaning the series appears to be close to stationary (autocorrelation function $\gamma(x) = I(x = 0)$).





b) The intercept and Lag2 appear to be statistically significant.

c)

Pred. / Actual	Down	Up
Down	54	48
Up	430	557

Roughly 56% accuracy. Most of the mistakes were overly optimistic, i.e. predicting the market would rise when in fact it fell. However, there were actually about 25% more ups than downs, which could be evidence that the logistic regression approach is a biased approach.

d)

Pred. / Actual	Down	Up
Down	9	5
Up	34	56

with 62.5% overall accuracy.

e) See .r file. Outcome is identical to logistic regression.

f)

Pred. / Actual	Down	Up
Down	0	0
Up	43	61

so evidently QDA never came up with a prediction of Down. It was accurate almost 59% of the time.

g)

Pred. / Actual	Down	Up
Down	21	29
Up	22	32

which is accurate almost 51% of the time.

h) LDA and logistic regression appear to perform best.

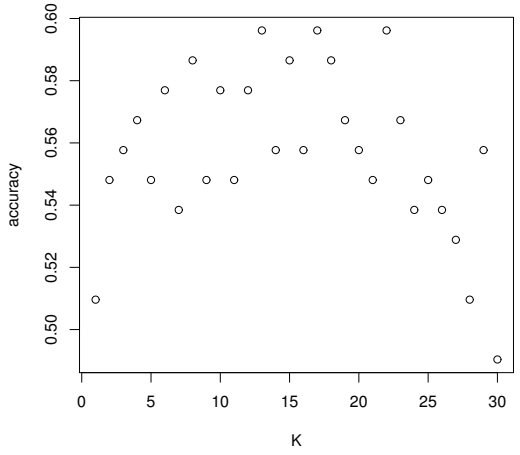
i) Plotting KNN accuracy for increasing values of K with only Lag2 as a predictor shows a high variance, with a maximum accuracy of maybe 58%.

The next most significant predictor is Lag1,

Method	Accuracy
Logistic	.587
LDA	.577
QDA	.558

If we suspect volume should make a difference, which I do, we can try the previous with Volume, Lag1, and Lag2 as predictors.

Method	Accuracy
Logistic	.605
LDA	.529
QDA	.462

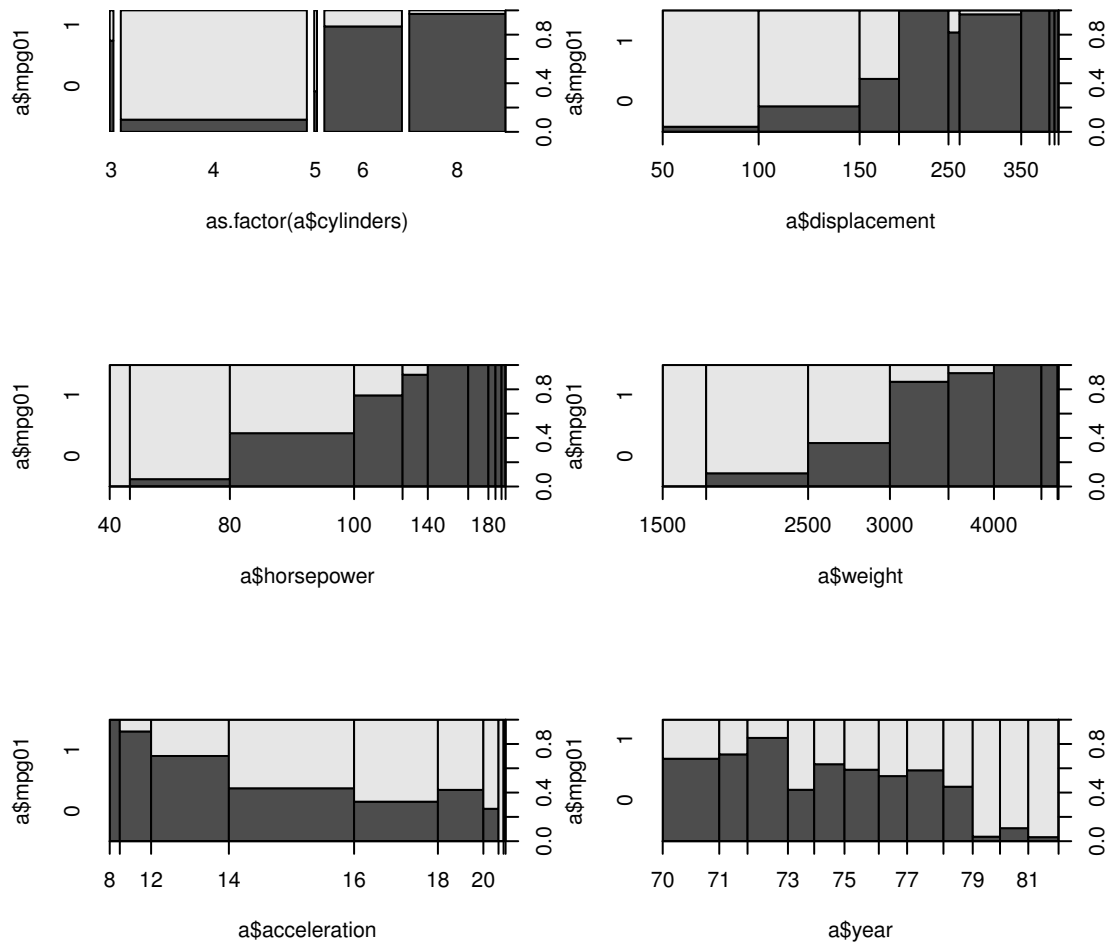


It appears that adding the volume term only helps the logistic regression. As a final test, we use only Lag2 and Volume*Lag2.

Method	Accuracy
Logistic	.625
LDA	.625
QDA	.529

The best method still is LDA or logistic regerssion on Lag2 alone, or Lag2 with Volume*Lag2.

11. b)



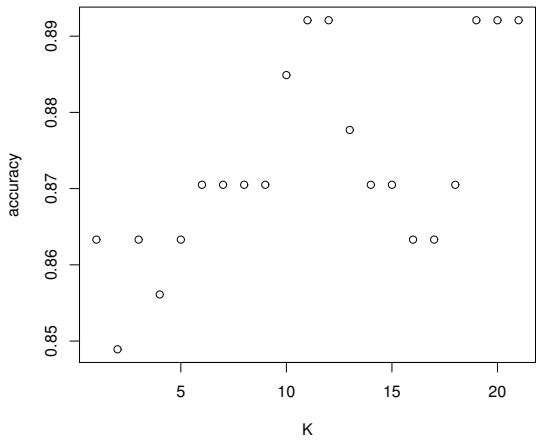
Number of cylinders, horsepower, and weight all appear to be exceptionally good indicators. Each has a positive correlation with mpg01, meaning more of any of the above is associated with a higher-than-median mpg.

d) LDA performs with about 1/10 test misclassification (error).

e) QDA performs with about 1/8 test misclassification.

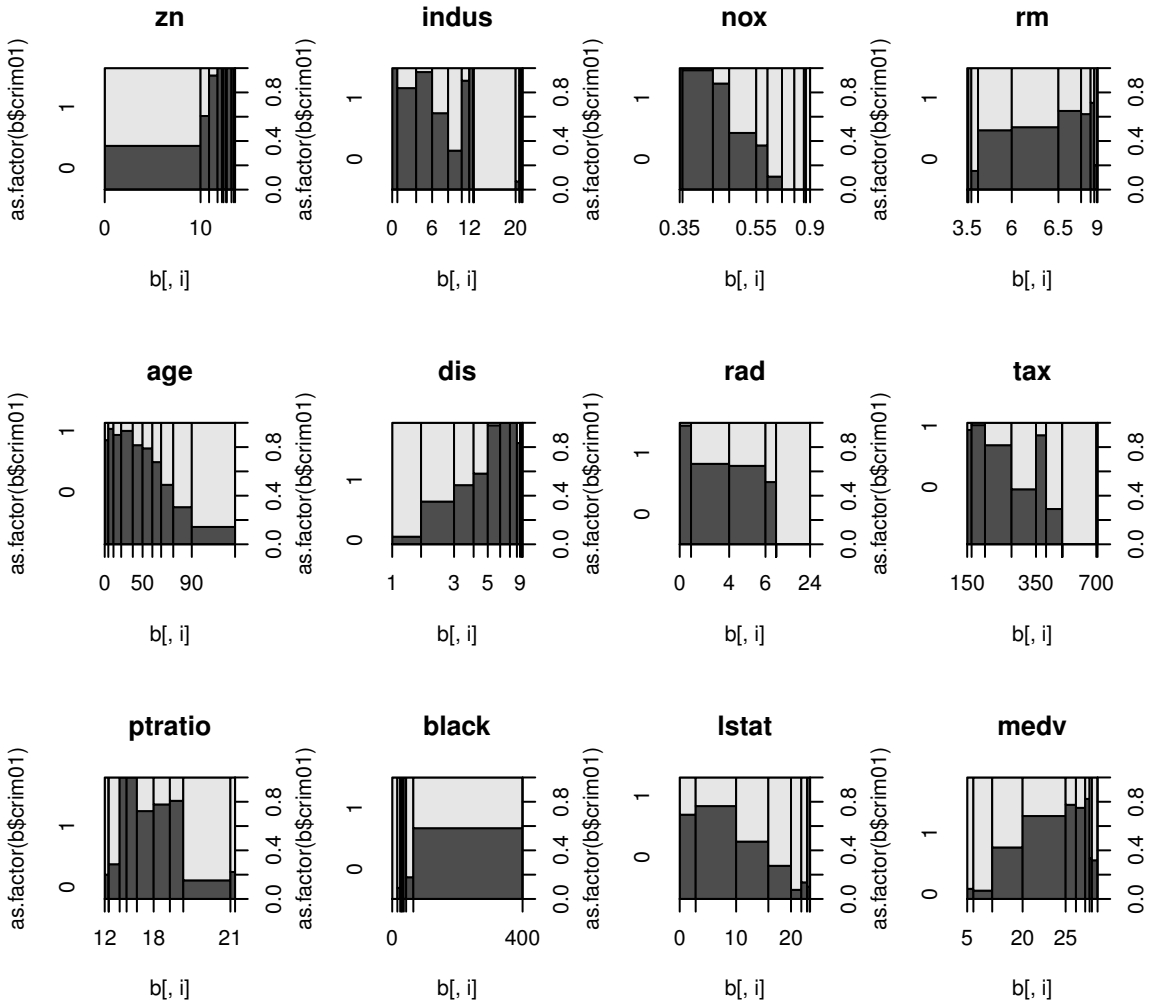
f) The logistic regression again performs as well as LDA.

g) $K = 11$ seem best. The test errors range from about .1 to .15, with $K = 11$ the least odd K (to avoid ties) among the best performing K.



12. See attached .r file

13. Call the relevant statistic crim01. Because chas is a binary variable, we wouldn't expect it to be a good predictor of crim01, which, as further evidenced with a plot (not pictured; see .r file), it isn't. Below are plots of crim01 (as a factor variable) vs. the rest of the covariates.



The strange scaling of the y-axis suggests that when plotting factors, the width of each bin is a linear function of the number of observations in that bin; the area of the rectangle remains the same as if the bins were equally spaced. This is an interesting way to do things; it can appear to show y-values more extreme than 0 and 1, but that speaks to the predictive power of a data point falling in a particular bin for a particular variable. Thus good indicators are ones with many observations falling into bins appearing to take values in the plots above close to (or especially more extreme than) 0 or 1. Thus the variables zn, indus, nox, dis, rad, tax, and ptratio will probably be relevant for classification.

Method	Accuracy
Logistic	.911
LDA	.829
1NN	.959
3NN	.966

It appears 3NN is the best method. Below is a plot of performance of odd K (to avoid ties) for KNN.