Homework 4 - Due 2/8/2017

Intermediate R

Complete the following course that you started last week: Datacamp R course. The website will automatically record your completion of the course, so don't worry about putting in a record of completion.

Poisson Regression

Poisson regression is used to model *count* data. It falls in a class of *generalized linear models* that extend linear regression and include things like logistic regression. These can generally be fit efficiently using the maximum likelihood estimation process that we described for linear regression, making them a good class of models for practical machine learning. R has a built in `glm` function which will fit them, very much like the `lm` function.

The *Poisson distribution* is a probability distribution on the non-negative integers

$$\mathbb{N} = 0, 1, 2, \ldots,$$

that has the following probability mass function:

$$P(n|\lambda) = \frac{e^{-\lambda}\lambda^n}{n!}$$

meaning that the probability assigned to the number n is given by the above expression and $\lambda > 0$ is a parameter. The probability assigned to n is meant to represent the probability of observing exactly n independent occurrences of an event that occurs at a given average rate over a fixed interval of time. For example, the number of cars passing by in the course of 10 minutes should follow a Poisson distribution.

1. Prove that this defines a probability distribution. The only thing to show here is that the total probability assigned to $\mathbb{N}$ is 1.

Hint: Remember the definitions of e^λ.

2. The λ parameter represents the *average rate* of events occurring. That means that λ should be the expected value of the random variable $f(x) = x$ on $\mathbb{N}$ with this Poisson distribution. Prove this fact.

Hints: Either remember the series defintion for e^λ or how to differentiate e^λ.

3. In fact, λ is also the variance of the above random variable. Prove this as well.

Hint: Another expression for the variance of a random variable X is $\mathbb{E}(X^2) - (\mathbb{E}(X))^2$.

4. Let Y be a target feature that we are seeking to predict and X an m-dimensional feature space. Then the model for Poisson regression is

$$\lambda(x) = e^{\beta \cdot x}$$

where y is an observed count that is sampled from $\lambda(x)$, and β is a vector in $\mathbb{R}^m$. Put another way, for each $x_i \in X$, we have some Poisson distribution determined by its rate $\lambda(x)$, and we assume the observed count data y_i is sampled from this distribution. Our prediction for this model is then $\lambda(x)$, as this represents the average value of the distribution defined by $\lambda(x)$ from problem 2.

Note: This is the same thing that we did in linear regression: for each x_i, we had a distribution $\beta \cdot x + N(0, \sigma)$ of possible y_i values, and our predicted value was the average: $\beta \cdot x$.

5. Given a training set $\{(x_i, y_i)\}$, write an expression for the log-likelihood as a function of β.

6. Write the log-likelihood function in R in the case where X is 1-dimensional and we have an intercept term. Do this as a function of the training data and $\beta = (\beta_0, \beta_1)$. That is, you should have a function

```
1_dim_poisson_log_lik <- function(beta, x, y){
...
}
```

that spits out a real number.

7. Use the R function `optim` to write a function that maximizes the above likelihood function over β in order to fit the regression model. This should be similar to last week's `one_dim_lm` function.

```
1_dim_poisson <- function(x,y){
...
}
```

It should spit out a named list with the estimated coefficients, predicted values, residuals.

Hint: Make sure you understand how the `optim` function works – it requires a real-valued function of the parameters to minimize, so in the `1_dim_poisson`, you will need to construct an intermediate likelihood function that is a function of just β.

8. Use the following data set to test your regression. It is a simulated data set where `num_awards` is the outcome variable and indicates the number of awards earned by students at a high school in a year, `math` is a continuous predictor variable and represents students' scores on their math final exam, and `prog` is a categorical predictor variable with three levels indicating the type of program in which the students were enrolled.

```
count_data <- read.csv("http://www.ats.ucla.edu/stat/data/poisson_sim.csv")
```

Use your function to fit a model `num_awards ~ math`. What do the coefficients tell you?

9. R has a built in `glm` function that extends `lm`, and can do Poisson regression, using the `family="poisson"` argument. Use this to fit a model that also includes `prog`. How does this compare? What do the coefficients tell you?