Problem 1 (multiclass logistic regression)

Part 1:

a. Write out the log likelihood L of data $(x_1, y_1), ..., (x_n, y_n)$ using an i.i.d. assumption.

First, we observe that while $p(y|x, w_1, ..., w_n)$ is a product of the likelihood of each w, the indicator effectively cancels out every $w_i, i \neq k$ as $z^{1(y=k)} = z^0 = 1$.

With that in mind, we can cast the joint probability of the data as:

$$L = \prod_{i=1}^{n} (\frac{e^{x_i^T w_i}}{\sum_{j=1}^{k} e^{x_i^T w_j}})^{1(y=i)}$$

Taking the natural log of this, we have:

$$lnL = \sum_{i=1}^{n} 1(y = i) ln(\frac{e^{x_i^T w_i}}{\sum_{j=1}^{k} e^{x_i^T w_j}})$$

$$lnL = \sum_{i=1}^{n} 1(y = i)(x_i^T w_i - ln \sum_{j}^{k} e^{x_i^T w_j})$$

b. Calculate $\nabla_{w_i} L$ and $\nabla^2_{w_i} L$.

Beginning with the log likelihood:

$$lnL = \sum_{i=1}^{n} 1(y = i)(x_i^T w_i - ln \sum_{j}^{k} e^{x_i^T w_j})$$

$$lnL'_{w_i} = \sum_{i=1}^{n} 1(y = i)(x_i - \frac{1}{\sum_{j}^{k} e^{x_i^T w_j}} (e^{x_i^T w_i}) x_i)$$

$$lnL'_{w_i} = \sum_{i=1}^{n} 1(y = i)(x_1)(1 - \frac{e^{x_i^T w_i}}{\sum_{j}^{k} e^{x_i^T w_j}})$$

The numerator $e^{x_i^T w_i}$ makes sense because we take the gradient with respect to w_i. Therefore, the terms of the summation are 0 for every $w_j, w_j \neq w_i$.

Part 2:

a. What is the joint likelihood of the data $(x1, ..., xN)$?

$$L_\lambda = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x!} e^{-\lambda}$$

b. Derive the maximum likelihood estimate $\hat{\lambda}_{ML}$ for λ.

$$lnL_\lambda = \sum_{i=1}^{n} x_i ln(\lambda) - ln(x!) - \lambda$$

$$lnL\prime_\lambda = \sum_{i=1}^{n} \frac{x_i}{\lambda} - 1 = 0$$

$$\frac{\sum x_i}{\lambda} = n$$

$$\frac{\sum x_i}{n} = \hat{\lambda}_{ML}$$

c. Explain why this maximum likelihood estimate makes intuitive sense.

Similarly to Part 1, the parameter λ of a poisson distribution is meant to represent the average rate of occurrence of the phenomenon being modeled. As such, the MLE of the parameter would be the average of all the values in the sample.

Problem 2 (Bayes rule)

a. Use Bayes rule to derive the posterior distribution of λ and identify the name of this distribution.

By Bayes rule, we know that:

$$P(\lambda|x) = \frac{P(x|\lambda)P(\lambda)}{P(x)}$$

Given that we are interested in λ, we can discard the denominator (which ultimately does not involve λ) and rewrite this equation as:

$$P(\lambda|x) \propto P(x|\lambda)P(\lambda)$$

We can then expand the right side of the equation into the joint distribution of $P(x, \lambda)$:

$$P(\lambda|x) \propto (\prod_{i=1}^{n} \frac{\lambda^{x_i}}{x!} e^{-\lambda}) \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} \lambda^{\alpha_1 - 1} e^{-\beta_1 \lambda}$$

$$\propto (\frac{\lambda^{\sum x_i}}{\prod x!} e^{-n\lambda}) \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} \lambda^{\alpha_1 - 1} e^{-\beta_1 \lambda}$$

$$\propto \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1) \prod x!} \lambda^{(\sum x_i + \alpha_1) - 1} e^{-\lambda(n + \beta_1)})$$

From this we have enough to learn the parameters of the posterior Gamma distribution: $\alpha_2 = \sum x_i + \alpha_1$ and $\beta_2 = n + \beta_1$. This result is still not the complete distribution, however: the fraction is not in the correct form for a Gamma, nor have we addressed the denominator which we dropped earlier. We know, however, that the final distribution must integrate to 1, which allows us to derive the correct values for the other terms (although we will not do so here).

b. What is the mean and variance of λ under this posterior? Discuss how this relates to your solution to Part 2 of Problem 1.

The posterior λ is $Gamma(\sum x_i + \alpha_1, n + \beta_1)$

As such, the mean of $\lambda = \frac{\alpha_2}{\beta_2} = \frac{\sum x_i + \alpha_1}{n + \beta_1}$.

The variance is $\frac{\alpha_2}{\beta_2^2} = \frac{\sum x_i + \alpha_1}{(n+\beta_1)^2}$.

In 1.2, we saw that: $\frac{\sum x_i}{n} = \hat{\lambda}_{ML}$.

There is a clear parallel to the Bayesian posterior mean: $\frac{\sum x_i + \alpha_1}{n + \beta_1} = E(\lambda)$, with the difference being the influence of the prior parameters α_1 and β_1. This relationship represents the Bayesian insight by which our posterior distribution is determined by an interaction between our prior beliefs (α_1, β_1) and our data $(\sum x_i, n)$.

Problem 3 (Linear regression)

Part 1

a. Print the numbers you obtain for the vector $\hat{w}_{ML}$. Using the labels of each dimension contained in the readme file, explain what the sign of each value in $\hat{w}_{ML}$ says about the relationship of the inputs to the output.

Dimension	Weight
Intercept	23.42352482
Num Cylinders	-0.5759929
Displacement	0.90318199
Horsepower	-0.12294213
Weight	-5.78808517
Acceleration	0.21887044
Model Year	2.77543049

Table 1: $\hat{w}_{ML}$

First, we note that when all dimensions are at their (centered) average, the intercept (average) for MPG is 23.42. Beyond that, the dimensions of Displacement, Acceleration, and Model Year are positively correlated with MPG, while Num Cylinders, Horsepower, and Weight, are negatively correlated.

Of these, we see that Weight is the most strongly negatively correlated with MPG, which is sensible as increased weight directly translates into increased energy necessary to move the car, reducing MPG. Interestingly, Model Year has by far the strongest positive correlation with MPG, suggesting that secular fuel efficiency improvements year over year play a bigger role in improving MPG than any other single factor.

b. Use the least squares solution to predict the outputs for each of the 20 testing examples. Repeat this process of randomly splitting into training and testing sets 1000 times. Each time, calculate the mean absolute error of the resulting predictions, MAE $= \frac{1}{20}\sum_{i=1}^{20} \mid y_{test} - y_{pred} \mid$. What is the mean and standard deviation of the MAE for these 1000 tests?

After running 1000 iterations, $\mu_{MAE} = 2.71242956285$, while $\sigma_{MAE} = 0.49817244422$

Part 2

a. In a table, print the mean and standard deviation of the RMSE as a function of p. Using these numbers argue for which value of p is the best.

As this table shows, setting $p = 3$ gives both the lowest RMSE and the lowest variance for any value of p.

p	Mean	Std
1	3.44030468834	0.672995051589
2	2.75919186824	0.623061482354
3	2.64085582004	0.602524538829
4	2.73039848277	0.697498334382

Table 2: RMSE by p

b. For each value of p, collect $y_{test} - y_{pred}$ for each test example. (Observe that this number can be negative, and there are 20 x 1000 in total.) Plot a histogram of these errors for each p.

c. For each p, use maximum likelihood to fit a univariate Gaussian to the 20,000 errors from Part 2(b). Describe how you calculated the maximum likelihood values for the mean and variance (this is a univariate case of what we did in class, so no need to re-derive it). What is the log likelihood of these empirical errors using the maximum likelihood values for the mean and variance? Show this as a function of p and discuss how this agrees/disagrees with your conclusion in Part 2(a). What assumptions are best satisfied by the optimal value of p using this approach?

To fit the Gaussian, we calculated both $\hat{\mu}_{ML}$ and $\hat{\sigma}^2_{ML}$ using the standard maximum-likelihood estimators for these parameters: $\frac{1}{n}\sum_{i=1}^{n} x_i$ and $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_{ML})^2$.

Looking at this table, $p = 3$ again appears to create the strongest model. The variance (although not the mean) of the error is lowest with this model. Additionally, the log likelihood of this error is the greatest, although by a small margin, compared to all other models. This supports our conclusions in 2(a), where we identified $p = 3$ as being the strongest model.

p	Mean	Var	Log Likelihood
1	0.0163194355852	3.50549766331	-53464.9206058
2	-0.0285981551706	2.82852246649	-49173.3602374
3	0.0409854631276	2.70840898169	-48305.4980812
4	0.0593496182852	2.81743994052	-49094.8436665

Table 3: Log Likelihood of error by p