Assignment 2

10/05/2020

Name: Yusuf Elnady
PID: 906280028

Q1.a

See the code.

Discussion about the results in KNN and Logistic Regression:

In KNN, the separators are not linear which enables us to fit the data easily and more flexibly than the case in Logistic Regression which uses the linear separability. In terms of Expressivity, KNN is more flexible and can express more complex relationships. Linear models are expected to perform better when the underlying function is linear and the true decision boundary is a linear one and less complex. The circumstances that restrict algorithm from the other to work well is that in KNN (non-linearity) we achieve higher accuracy when number of neighbors is small. The key is that in a large number of nearest neighbors, the data can overfit which is not the case in the logistic regression.

For our dataset, the logistic regression gives a higher accuracy than KNN: we are only classifying two classes, although the underlying function is not linear, but using a linear separator was excellent to give a higher accuracy. If the number of classes were 15 or so, then the KNN will be better to use and give a higher accuracy.

Q2.a

The likelihood function for a data set D is given by $p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}$. The maximum likelihood estimate of θ is the value that maximizes the likelihood function. So, we will start by calculating the derivative $\frac{\mathrm{d}p(\mathcal{D}|\theta)}{\mathrm{d}\theta}$ then equalling it to 0, and see the value of $\hat{\theta}$ in this case.

First we take the log of the likelihood function:

$$\ln(p(\mathcal{D}|\theta)) = N_1 \ln(\theta) + N_0 \ln(1-\theta)$$

By taking the derivative:

$$\frac{\mathrm{d}p(\mathcal{D}|\theta)}{\mathrm{d}\theta} = \frac{N_1}{\theta} - \frac{N_0}{1-\theta}$$

The next step is to choose the value of θ that maximizes the log of the likelihood function:

$$\frac{\mathrm{d}p(\mathcal{D}|\theta)}{\mathrm{d}\theta} = 0 = \frac{N_1}{\theta} - \frac{N_0}{1-\theta}$$
$$0 = \theta(N_0) - (1-\theta)N_1 = \theta N_0 - N_1 + \theta N_1 = \theta(N_0 + N_1) - N1$$
$$\hat{\theta} = \frac{N_1}{N_0 + N_1}$$

To make sure this is the maximum value, not the minimum, we can take the second derivative $\frac{\mathrm{d}^2 p(\mathcal{D}|\theta)}{\mathrm{d}\theta^2}$ and by substituting the $\hat{\theta}$ we got we will found that the value will be lower than 0. which mean this $\hat{\theta} = \frac{N_1}{N_0+N_1}$ is the solution for the maximum likelihood and we have got it from eq:bernoulli to eq:mle.

Q2.b

Given $p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$, we see that $p(\mathcal{D})$ will not change when $\mathcal{D}$ is fixed, therefore we can get the value of $\hat{\theta}$ that maximizes the posterior likelihood $p(\theta|\mathcal{D})$ by solving only for the maximum point of $p(\mathcal{D}|\theta)p(\theta)$.

Note: we have also previously calculated the maximum for $p(\mathcal{D}|\theta)$.

The first step is to take the derivative:

$$\frac{\mathrm{d}p(\mathcal{D}|\theta)p(\theta)}{\mathrm{d}\theta} = \frac{\mathrm{d}(\theta^{N_1+\alpha}(1-\theta)^{N_0+\alpha})}{\mathrm{d}\theta}$$
$$= (\alpha + N_1)\theta^{N_1+\alpha-1}(1-\theta)^{N_0} - (\alpha + N_0)\theta^{N_1+\alpha}(1-\theta)^{N_0+\alpha-1}$$
$$= (N_1 + \alpha)(1 - \theta\frac{N_1 + N_0 + 2\alpha}{N_1 + \alpha})\theta^{N_1-1}(1-\theta)^{N_0-1}$$

Then by making the derivative equal to zero, we get:

$$0 = (N_1 + \alpha)(1 - \theta\frac{N_1 + N_0 + 2\alpha}{N_1 + \alpha})\theta^{N_1-1}(1-\theta)^{N_0-1}$$
$$\frac{N_1 + \alpha}{\theta} - \frac{N_0 + \alpha}{1-\theta} = (N_1 + \alpha)(1-\theta) = \theta(N_0 + \alpha)$$
$$= (N_1 + \theta) - \alpha(N_0 + N_1 + 2\alpha)$$

Therefore $\hat{\theta} = \frac{N_1+\alpha}{N_0+N_1+2\alpha}$ is the maximum posterior mean estimate.

To make sure this is the maximum value, not the minimum, we can take the second derivative $\frac{\mathrm{d}p(\mathcal{D}|\theta)p(\theta)}{\mathrm{d}\theta^2}$ and by substituting the $\hat{\theta}$ we got we will found that the value will be lower than 0 because at $\hat{\theta} = \frac{N_1+\alpha}{N_0+N_1+2\alpha}$ we find that

$$\frac{\mathrm{d}p^2(\mathcal{D}|\theta)p(\theta)}{\mathrm{d}\theta^2} = \frac{-(N_1+\alpha)}{\alpha^2} + \frac{-(N_0+\alpha)}{(1-\alpha)^2} < 0$$

.

Q3.a

$P(C_0|x)$: Probability of being classified as having benign lung cancer given the observed symptoms x.

$P(x|C_0)$: Probability of having the observed symptoms given that the classification is having benign lung cancer.

$P(C_0)$: Probability of having benign lung cancer.

$P(C_1|x)$: Probability of being classified as having malign lung cancer given the observed symptoms x.

$P(x|C_1)$: Probability of having the observed symptoms given that the classification is having malign lung cancer.

$P(C_1)$: Probability of having malign lung cancer.

Q3.b

We have $P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}$ so if $k = 0$ then we have $P(C_0|x) = \frac{P(x|C_0)P(C_0)}{P(x)}$.

We also know that in conditional probabilities we have

$$P(C_0|x) + P(C_1|x) = 1$$

To calculate for $P(C_0|x)$, we can put it in this way:

$$P(C_0|x) = 1 - P(C_1|x) = 1 - \frac{1}{1 + \exp^{-w^T x}}$$

$$P(C_0|x) = \frac{\exp^{-w^T x}}{1 + \exp^{-w^T x}}$$

Now, to calculate $\log \frac{P(C_1|x)}{P(C_0|x)}$, we can just substitute their values and simplify:

$$\log \frac{P(C_1|x)}{P(C_0|x)} = \log \frac{(\frac{1}{1+\exp^{-w^T x}})}{(\frac{\exp^{-w^T x}}{1+\exp^{-w^T x}})} = \log(\frac{1}{\exp^{-w^T x}})$$

$$\log \frac{P(C_1|x)}{P(C_0|x)} = \log(\exp^{+w^T x}) = w^T x$$

Q4.a

A discrete random variable X follows a Poison distribution with parameter λ if $Pr(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$. Since each trial is independnet, the probability is just simply the product of all entries. We can start by calculating the **likelihood function** as

$$P(G|\lambda) = \prod_{i=0}^{n} P(G_i) = \prod_{i=0}^{n} \frac{\lambda^{G_i} e^{-\lambda}}{G_i!}$$

$$P(G|\lambda) = \frac{\lambda^{\sum_i^n G_i} e^{-n\lambda}}{\prod_{i=0}^n G_i!}$$

By taking the natural log for both sides to calculate **the log-likelihood function of G given** λ. We get:

$$\ln(P(G|\lambda) = \ln(\frac{\lambda^{\sum_i^n G_i} e^{-n\lambda}}{\prod_{i=0}^n G_i!}) = \ln(\lambda^{\sum_i^n G_i}) + \ln(e^{-n\lambda}) - \ln(\prod_{i=0}^n G_i!)$$

$$\ln(P(G|\lambda)) = \sum_i^n G_i \ln(\lambda) - \lambda n - \ln(\prod_{i=0}^n G_i!)$$

Q4.b

To get the MLE from the likelihood function, we take its derivative with respect to λ and then set it to zero, and finally solve for λ.

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln(P(G|\lambda)) = \frac{\mathrm{d}}{\mathrm{d}\lambda}(\ln(\lambda^{\sum_i^n G_i}) + \ln(e^{-n\lambda}) - \ln(\prod_{i=0}^n G_i!))$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln(P(G|\lambda)) = \frac{\sum_{i=1}^n G_i}{\lambda} - n - 0 = 0$$

After setting the derivative to zero, we get the following λ of the MLE:

$$\lambda = \frac{\sum_{i=1}^n G_i}{n}$$

Q4.c

To compute the MLE for λ using the observed G, it will be:

$$\lambda = \frac{4+1+3+5+5+1+3+8}{8} = \frac{15}{4} = 3.75$$