Homework 1: k-Nearest Neighbor classification and Statistical Estimation

#1 Statistical Estimation

1) $D = \{x_1, x_2, \ldots, x_N\}$ from N pois($x = x_i|\lambda$)

$P(x=x) = \frac{e^{-\lambda}\lambda^x}{\lfloor x}$, $n = 0, 1, 2 \cdots$

$$P(D|\lambda) = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_{i=1}^{n} \lfloor x_i}$$

$$\log P(D|\lambda) = -n\lambda + \sum x_i \log\lambda - \log\left[\prod_{i=1}^{n} \lfloor x_i\right]$$

$$\frac{\partial}{\partial\lambda}\log P(D|\lambda) = -n + \frac{\sum x_i}{\lambda} - 0 = 0$$

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

$$\Rightarrow \hat{\lambda} = \bar{x} \rightarrow \lambda \text{ is the average \# of occurrence}$$

2) ① So, $f(x_i|\lambda) = e^{-\lambda}\frac{\lambda^{x_i}}{\lfloor x_i}$, $x_i = 0, 1, 2 \cdots$

$$\Rightarrow f(\lambda) = \frac{e^{-\lambda\beta}\lambda^{\alpha-1}}{\lfloor(\alpha)}\beta^{\alpha}, \quad \lambda > 0$$

So, $f(x_i, \lambda) = \prod_{i=1}^{N} f(x_i|\lambda)\, p(\lambda)$

$$= e^{-N\lambda}\frac{\lambda^{\sum_{i=1}^{N} x_i}}{\prod_{i=1}^{N}\lfloor x_i}\cdot\frac{e^{-\lambda\beta}\lambda^{\alpha-1}}{\lfloor(\alpha)}\beta^{\alpha}$$

$$\Rightarrow f(x) = \int_0^{\infty} f(x, \lambda)\, d\lambda$$

$$= \int_0^{\infty} \frac{\beta^{\alpha}}{\prod_{i=1}^{N}\lfloor x_i \lfloor(\alpha)} e^{-N\lambda} e^{-\lambda\beta}\lambda^{\sum_{i=1}^{N} x_i}\lambda^{\alpha-1}\, d\lambda$$

$$= \frac{\beta^{\alpha}}{\prod_{i=1}^{N}\lfloor x_i \lfloor(\alpha)} \int_0^{\infty} e^{-\lambda(N+\beta)}\lambda^{\sum_{i=1}^{N} x_i+\alpha-1}\, d\lambda$$

Here, $\Rightarrow p(x|D) = \frac{f(x)\cdot\lambda}{f(x)}$

② $$\Rightarrow \frac{\cancel{\beta^{\alpha}}\cancel{e^{-N\lambda}}\, e^{-\lambda(N+\beta)}\lambda^{\sum_{i=1}^{N} x_i+\alpha-1}}{\cancel{\prod_{i=1}^{N} x_i \lfloor(\alpha)}\;\frac{\Gamma\left(\sum_{i=1}^{N} x_i+\alpha\right)}{(\lambda+\beta)^{\sum_{i=1}^{N} x_i+\alpha}}}$$

$$\Rightarrow \frac{1}{\Gamma\left(\sum_{i=1}^{N} x_i+\alpha\right)} e^{-\lambda(N+\beta)}\lambda^{\sum_{i=1}^{N} x_i+\alpha-1}(N+\beta)^{\sum_{i=1}^{N} x_i+\alpha}$$

$$\ln p(x|D) \Rightarrow -\ln\left[\Gamma\left(\sum_{i=1}^{N} x_i+\alpha\right)\right] - \lambda(N+\beta) + \left(\sum_{i=1}^{N} x_i+\alpha-1\right)\ln\lambda + \left(\sum_{i=1}^{N} x_i+\alpha\right)\ln(\lambda+\beta)$$

$$= C - \lambda(N+\beta) + \lambda\ln\lambda\left[\sum_{i=1}^{N} x_i+\alpha-1\right]$$

$$= \frac{d}{d\lambda}\cdot p(\lambda|D) = 0 + \frac{d}{d\lambda}\left[-\lambda(N+\beta)\right] + \frac{d}{d\lambda}\left[+\lambda\left[\sum_{i=1}^{N} x_i+\alpha-1\right]\right]$$

$$= -(N+\beta) + \frac{\sum_{i=1}^{N} x_i+\alpha-1}{\lambda}$$

So, $\frac{d}{d\lambda}\ln p(\lambda|D) = 0$

$$\Rightarrow -(N+\beta) + \frac{\left(\sum_{i=1}^{N} x_i+\alpha-1\right)}{\lambda} = 0$$

$$\Rightarrow \left(\sum_{i=1}^{N} x_i+\alpha-1\right) = \lambda(N+\beta)$$

$$\Rightarrow \lambda = \frac{\sum_{i=1}^{N} x_i+\alpha-1}{N+\beta} = \hat{\lambda}$$

③ MAP estimator means maximum of a posterior estimator. for example, here in this problem the maximum value of a parameter λ which maximisizes the posterior probability distribution of $P(\lambda|D)$. Therefore, $\lambda = \frac{\sum_{i=1}^{N} x_i+\alpha-1}{N+\beta}$ is a MAP estimator of 1

3) Let $D|\lambda = Y_1, Y_2 \dots Y_n | \lambda \overset{iid}{\sim} \text{poisson}(\lambda)$ and $\lambda \sim \text{Gamma}(\alpha, \beta)$

The conditional joint density of $y_1, y_2 \dots y_n|\lambda$ is $f(y_1, y_2 \dots y_n|\lambda) = \prod_{i=1}^{n} \left(\frac{e^{-\lambda}\lambda^{y_i}}{y_i!}\right) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!}$

The prior is $\pi(\lambda) = \frac{e^{-\beta\lambda}\lambda^{\alpha-1}\beta^{\alpha}}{\Gamma\alpha} \quad \lambda > 0$

The posterior distribution $\Rightarrow \pi(\lambda|D) = \pi(\lambda|y_1 \dots y_n) = \frac{f(y_1, y_2 \dots y_n|\lambda)\,\pi(\lambda)}{\int_0^{\infty} f(y_1 \dots y_n|\lambda)\,\pi(\lambda)\,d\lambda}$ } This is independent of λ, meaning

$\pi(\lambda|D) \propto f(y_1, y_2 \dots y_n|\lambda)\,\pi(\lambda)$

$\Rightarrow \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} y_i}}{\left(\prod_{i=1}^{n} y_i!\right)} \times \frac{e^{-\beta\lambda}\lambda^{\alpha-1}\beta^{\alpha}}{\Gamma\alpha}$

$\propto e^{-\lambda(\beta+n)}\,\lambda^{\left(\sum_{i=1}^{n} y_i + \alpha - 1\right)}$

$\lambda|D \sim \text{gamma}(d_p, \beta_p)$ where $\alpha_p = \sum_{i=1}^{n} y_i + \alpha - 1$

$\beta_p = \beta + n$

Hence the gamma distribution is a conjugate prior to the poisson as seen.

#2 k-Nearest Neighbor (kNN)

4) One-hot encoding techiques would make it hard when calculating euclidean distance. The values are all 1s and 0s. since it is all zeros and ones this can result in a zero value for each label. (This can be handled by setting Unknown to ignore)

↳ example: [1,0,0], [0,1,0], [0,0,1]]

Catergorical varibles as ordinals would potential provide smaller matrixes when solving for Euclidian distance in knn.

↳ example: [1], [2], [3]]

5) 100 % of the training data has an income > 50k. This can affect the model because we only have data from higher income individuals. This means if it was trying to make a perdiction for someone who makes <= 50k it would be very lost because it has no data to go off of. 70 % accuracy is good depending on what the model is testing for, weather it is good or not will vary b/w situations. This model has 12 dimensions

6) Let $x = (x_1, x_2, \dots x_d)^T$ be a vector. zero vector $\Rightarrow x = (0, 0, \dots 0)^T$ so, The Euclidean distance b/w x and the zero vector $\Rightarrow \sqrt{(x_1-0)^2 + (x_2-0)^2 + \dots + (x_d-0)^2} \Rightarrow \|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$

Let $z = (z_1, z_2, \dots z_d)^T$ The Eucliden distance b/w x and z vector $\Rightarrow \sqrt{(x_1-z_1)^2 + (x_2-z_2)^2 + \dots + (x_d-z_d)^2}$

$\Rightarrow \|x-z\|_2 = \sqrt{\sum_{i=1}^{d} (x_i-z_i)^2}$ $\Rightarrow$ implies that, distance b/w x and z can be written as an ℓ_2 norm

7) coding

8) coding

9) The best k I observed is 9 or 7 because overall they had the highest training accuracy and validation accuracy when k=1 the training error is close to 0 % when you choose a test sample closets to your training set. k=1 will not be zero when the test sample is not close to your training set.

A trend you may observe is the higher k is the lower the training accuracy but the higher the validation accuracy is.

Underfitting occurs when the training data does not quite fit the trends of real data, meaning our model does not fit well enough. This can kinda be seen by lower k's, like k of 1 because validation its 78.72 % but it does well with traning, getting a 98.74, so its not the best example

An example of overfitting can be seen when k is too large and it causes training to go awry. This can be seen with k = 8000 and train acc = 75.41 %.

10)

#3 Debriefing

1) 3.5 days (yes I barely started it the day before it was due and no that will not be happening again)
2) difficult, I was literally brought to tears
3) mostly alone, was gonna ask TAs but I realized there are no office hours on Thursdays and I missed Wednesday but fridays was helpful
4) I think at first 30% but after struggling through maybe like 85% now
5) sorry its so slow, I did not want to code anymore