CS434_HW1-4

Lyon Kee

October 10, 2023

1 Statistical Estimation [10pts]

Given:

$$Pois(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \forall x \in 0, 1, 2, ... \quad \lambda \geq 0 \tag{1}$$

1.1 Q1 Maximum Likelihood Estimation of λ

1.1.1 Write out the log-likelihood function $\log(\mathbf{P}(\mathbf{D}|\lambda))$

$$\begin{aligned}
P(D|\lambda) &= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\
\log P(D|\lambda) &= \sum_{i=1}^{n} \log \frac{\lambda^{x_i} e^{-\lambda}}{x!} \\
&= \sum_{i=1}^{n} \big(\log(\lambda^{x_i} e^{-\lambda}) - \log(x_i!)\big) \\
&= \sum_{i=1}^{n} \big(\log(\lambda^{x_i}) + \log(e^{-\lambda}) - \log(x_i!)\big) \\
&= \sum_{i=1}^{n} (x_i \log(\lambda) - \lambda - \log(x_i!))
\end{aligned}$$

1.1.2 Take the derivative of the log-likelihood with respect to the parameter λ

$$\begin{aligned}\frac{P(D|\lambda)}{\delta\lambda} &= \frac{\sum_{i=1}^{n}\left(x_i\log(\lambda) - \lambda - \log(x_i!)\right)}{\delta\lambda}\\ &= \sum_{i=1}^{n}\left(\frac{x_i\log(\lambda)}{\delta\lambda} - \frac{\lambda}{\delta\lambda} - \frac{\log(x_i!)}{\delta\lambda}\right)\\ &= \sum_{i=1}^{n}\left(\frac{x}{\lambda} - 1\right)\end{aligned}$$

1.1.3 Set the derivative equal to zero and solve for λ – call this maximizing value $\hat{\lambda}_{MLE}$

$$\begin{aligned}\frac{P(D|\lambda)}{\delta\lambda} &= 0\\ \sum_{i=1}^{n}\left(\frac{x_i}{\lambda} - 1\right) &= 0\\ \frac{\sum_{i=1}^{n} x_i}{\lambda} - N &= 0\\ \frac{\sum_{i=1}^{n} x_i}{\lambda} &= N\\ \sum_{i=1}^{n} x_i &= N\lambda\\ \hat{\lambda}_{MLE} &= \frac{\sum_{i=1}^{n} x_i}{N}\end{aligned}$$

1.2 Q2 Maximum A Posteriori Estimate of λ with a Gamma Prior [4pts]

Given:

$$Gamma(\Lambda = \lambda; \alpha, \beta) = \frac{\beta^{\alpha}\lambda^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)} \quad \forall\lambda \geq 0 \quad \alpha, \beta \geq 0 \tag{2}$$

1.2.1 Write out the log-posterior $\log P(\lambda|D) \propto \log P(D|\lambda) + \log P(\lambda)$

$$
\begin{aligned}
\log P(\lambda|D) &\propto \log P(D|\lambda) + \log P(\lambda) \\
&\propto \sum_{i=1}^{n} (x_i \log(\lambda) - \lambda - \log(x_i!)) + \log \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} \\
&\propto \sum_{i=1}^{n} (x_i \log(\lambda) - \lambda - \log(x_i!)) + \log \beta^\alpha + \log \lambda^{\alpha-1} + \log e^{-\beta\lambda} - \log \Gamma(\alpha) \\
&\propto \sum_{i=1}^{n} (x_i \log(\lambda) - \lambda - \log(x_i!)) + \alpha \log \beta + (\alpha - 1) \log \lambda + (-\beta\lambda) \log e - \log \Gamma(\alpha) \\
&\propto \sum_{i=1}^{n} (x_i \log(\lambda) - \lambda - \log(x_i!)) + \alpha \log \beta + (\alpha - 1) \log \lambda - \beta\lambda - \log \Gamma(\alpha)
\end{aligned}
$$

1.2.2 Take the derivative of $\log P(D|\lambda) + \log P(\lambda)$ with respect to the parameter λ

$$
\begin{aligned}
\frac{\log P(\lambda|D)}{\delta\lambda} &\propto \frac{\log P(D|\lambda) + \log P(\lambda)}{\delta\lambda} \\
&\propto \frac{\sum_{i=1}^{n} (x_i \log(\lambda) - \lambda - \log(x_i!)) + \alpha \log \beta + (\alpha - 1) \log \lambda - \beta\lambda - \log \Gamma(\alpha)}{\delta\lambda} \\
&\propto \frac{\sum_{i=1}^{n} (x_i \log(\lambda) - \lambda - \log(x_i!)) + \alpha \log \beta + \alpha \log \lambda - \log \lambda - \beta\lambda - \log \Gamma(\alpha)}{\delta\lambda} \\
&\propto \sum_{i=1}^{n} \left(\frac{x_i}{\lambda} - 1 - 0\right) + 0 + \frac{\alpha}{\lambda} - \frac{1}{\lambda} - \beta - 0 \\
&\propto \sum_{i=1}^{n} \left(\frac{x_i}{\lambda}\right) - n + \frac{\alpha}{\lambda} - \frac{1}{\lambda} - \beta
\end{aligned}
$$

1.2.3 Set the derivative equal to zero and solve for λ – call this maximizing value $\hat{\lambda}_{MAP}$

$$\begin{aligned}
\frac{\log P(\lambda|D)}{\delta\lambda} &= 0 \\
\sum_{i=1}^{n}\left(\frac{x_i}{\lambda}\right) - n + \frac{\alpha}{\lambda} - \frac{1}{\lambda} - \beta &= 0 \\
\frac{\sum_{i=1}^{n}(x_i)}{\lambda} + \frac{\alpha}{\lambda} - \frac{1}{\lambda} &= \beta + n \\
\frac{\sum_{i=1}^{n}(x_i) + \alpha - 1}{(\beta + n)} &= \lambda \\
\hat{\lambda}_{MAP} = \frac{\sum_{i=1}^{n}(x_i) + \alpha - 1}{(\beta + n)}
\end{aligned}$$

1.3 Q3 Deriving the Posterior of a Poisson-Gamma Model [2pt]

$$\begin{aligned}
&P(\lambda|D) \propto Gamma(\lambda; \alpha_P, \beta_P) \\
&P(\lambda|D) \propto P(D|\lambda)P(\lambda) \\
&P(\lambda|D) \propto \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \cdot \frac{\beta_P^{\alpha_P} \lambda^{\alpha_P - 1} e^{-\beta_P \lambda}}{\Gamma(\alpha_P)} \\
&P(\lambda|D) \propto \prod_{i=1}^{n} \lambda^{x_i} e^{-\lambda} \cdot \frac{\beta_P^{\alpha_P} \lambda^{\alpha_P - 1} e^{-\beta_P \lambda}}{\Gamma(\alpha_P)} \\
&P(\lambda|D) \propto \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda} \cdot \frac{\beta_P^{\alpha_P} \lambda^{\alpha_P - 1} e^{-\beta_P \lambda}}{\Gamma(\alpha_P)} \\
&P(\lambda|D) \propto \frac{\beta_P^{\alpha_P} \lambda^{\alpha_P - 1 + \sum_{i=1}^{n} x_i} e^{-\beta_P \lambda - n\lambda}}{\Gamma(\alpha_P)} \\
&P(\lambda|D) \propto \frac{\beta_P^{\alpha_P} \lambda^{(\alpha_P + \sum_{i=1}^{n} x_i) - 1} e^{-\lambda(\beta_P + n)}}{\Gamma(\alpha_P)}
\end{aligned}$$

$\therefore$we observe that Gamma is proportional with:

$$\alpha_p = \alpha_P + \sum_{i=1}^{n} x_i$$

$$\beta_p = \beta_P + n$$

2 k-Nearest Neighbor (kNN) [50pts]

2.1 Q4 Encodings and Distance [2pt]

The original field will have a large value of Euclidean distance, however when we perform one hot encoding or ”binarization” we observe that the Euclidean distances would differ by $\sqrt{1^2+1^2} = \sqrt{2}$. This would then make it so that the order of private to state-gov would not introduce big Euclidean distance for no reason.

2.2 Q5 Looking at Data [3pt]

$1967/8000 = 12.09\%$ has an income $> 50k$. This will most likely make it so that the model will predict a 0 for income more often than 1 for income despite the features telling otherwise. If a model as 70% accuracy, it is most likely a bad model because if we assume this sample as the population, we are able to predict a 0 for income all the time and achieve an 87.91% accuracy.
age: 1
education: 1
capital-gain: 1
capital-loss: 1
hours-per-week: 1
workclass: 7
martial-status: 7
occupation: 14
relationship: 6
race: 5
sex: 1 native-country: 40

2.3 Q6 Norms and Distances [2pt]

$$L_2(x) = \sqrt{\sum_{i=1}^{d} x_i^2}$$

$$L_2(x,z) = \sqrt{\sum_{i=1}^{d} (x_i - z)^2} \quad \text{Given } z = z_i \quad \forall z_i$$

L_2 norm can be calculated with np.linalg.norm by obtaining the difference between each axis by performing vector subtraction. Thus generating the code np.linalg.norm(x-z)

2.4 Q7 Implement kNN Classifier [10pts]

2.5 Q8 Implement k-fold Cross Validation [8pts]

2.6 Q9 Hyperparameter Search [15pt]

2.7 What is the best number of neighbors (k) you observe?

Out of the range of K provided, k = 30 was the best k. But with further investigation and other k values, I found k = 55 as the best k.

2.8 When k = 1, is training error 0%? Why or why not?

When k = 1, we do not observe 0% training error. This is because points equidistant from the target point will both be the closest K, but also the fact that there might be two points together with different labels thus generating error because both points would result in the same label prediction.

2.9 What trends (train and cross-validation accuracy rate) do you observe with increasing k?

training decreases as k goes further away from 1 because this gets the model away from overfitting to the training set. cross-validation rate and accuracy are low because it is overfitted. As K increases, there is a sweet spot until underfitting occurs. As K becomes large enough for overfitting, we observe validation rate and accuracy because it is so underfitted that it always predicts the majority label.

2.10 How do they relate to underfitting and overfitting?

As stated above, as K goes from low to high, we observe overfitting and then underfitting. There is a sweet spot in the middle where it is not underfit or overfit but we will not observe high accuracy on the training set because it is not overfitted.

3 Debriefing (required in your report)

3.1 Approximately how many hours did you spend on this assignment?

around 15 hours with 3 hours of sleep. :)

3.2 Would you rate it as easy, moderate, or difficult?

the math part is slightly confusing because I didn't know how a distribution affected prior and likelihood when calculating posterior, that was very confusing. Another issue I had was with how the product of all works as this had confused me to where I thought there are no solutions.

3.3 Did you work on it mostly alone or did you discuss the problems with others?

I worked on it alone.

3.4 How deeply do you feel you understand the material it covers (0%–100%)?

80%

3.5 Any other comments?

Nope