Andrea Tongsak
HOMEWORK 1

Andrea Tangsak CS 434 Due: Wed Oct 12
+3 days Extension Granted due to Hospital Visit

1. Statistical Estimation [10 pts]

Maximum Likelihood Estimation of λ

Q1

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

$D = \{x_1, x_2, \ldots, x_n\}$ i.i.d $\text{Pois}(X=x;\lambda)$

Let $\hat{\lambda}$ be the MLE for λ

$$f(x_i;\lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

$$\prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = L(\lambda) \text{ or likelihood of } \lambda$$

$$\log P(D|\lambda) = L(\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n}(x_i!)}$$

1. Take log-likelihood function $\log P(D|\lambda)$

$$\log L(\lambda) = \left(\sum_{i=1}^{n} x_i\right)\log\lambda - n\lambda - \sum_{i=1}^{n}\log(x_i!)$$

Take derivative of the log-likelyhood with respect to λ

$$\frac{d \log L(\lambda)}{d\lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n$$

Set derivative equal to $\emptyset$ and solve for λ (call $\hat{\lambda}$)

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

$$\hat{\lambda}_{LME} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Q2 $P(\lambda | D) \propto P(D|\lambda) P(\lambda)$

$$\propto \lambda^n e^{-\lambda \Sigma_i x_i} \lambda^{\alpha - 1} e^{-\beta\lambda}$$

$$P(\lambda | D) \propto e^{-\lambda(\Sigma_i x_i - \beta)} \lambda^{n+\alpha-1}$$

so $P(\lambda | D) \propto \text{Gamma}(\alpha + n, \Sigma_i x_i + \beta)$

Let $\hat{\lambda}_{MAP}$ = the maximum a posteriori estimator (MAP) as a function of α, β

log posteriori is

$$\log P(\lambda | D) \propto -\lambda\left(\sum_i x_i + \beta\right) + (n + 2 - 1) \log \lambda$$

set the derivative to 0:

$$\frac{d}{d\lambda} \log P(\lambda | D) \propto -\sum_i x_i - \beta + \frac{(n+2-1)}{\lambda} = \emptyset$$

$$\frac{(n + \alpha - 1)}{\lambda} = \sum_i x_i + \beta$$

Thus $\hat{\lambda}_{MAP} = \lambda = \dfrac{n + \alpha - 1}{\Sigma_i x_i + \beta}$

Q3

$D = \{x_1, x_2, \dots, x_n\}$ i.i.d. $Pois(X = x; \lambda)$

Likelihood of λ $\quad L(\lambda|x) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\prod_{i=1}^{n}(x_i!)}$

Prior $\quad p(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \lambda > 0$

$\Rightarrow$ Posterior by Bayes Rule

$$P(\lambda|D) = \frac{P(D|\lambda) \cdot P(\lambda)}{P(D)}$$

$\rightarrow$ No λ here, we can ignore it.

So we can rewrite

$$\frac{P(D(\lambda) \cdot P(\lambda))}{P(D)} \propto P(D|\lambda) \cdot P(\lambda)$$

$$P(D|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{x_1+x_2+\dots+x_n} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

- no λ here, we can ignore

$$P(D|\lambda) \propto \lambda^{x_1+x_2+\dots+x_n} \cdot e^{-n\lambda} = \lambda^{nX} \cdot e^{-n\lambda}$$

she $\sum_{i=1}^{n} x_i = nX$

$$P(\lambda|D) \propto P(D|\lambda) \cdot P(\lambda) = \lambda^{nX} e^{-n\lambda} \cdot \underbrace{\lambda^{\alpha-1} e^{-\beta\lambda}}_{\text{Prior}}$$

$$= \lambda^{nX+\alpha-1} e^{-(\beta+n)\lambda}$$

} This is actually gamma $(nX+\alpha, \beta+n)$

Since $\lambda \propto$ gamma $(\alpha, \beta$

$x_i \propto Pois(\lambda)$

Then Gamma distribution is a conjugate prior to Poisson

2. k-Nearest Neighbor (kNN)

Q4. One hot encoding technique = transforming each possible value into binary attribute.

Difference between encoding workclass as binary vs. ordinal

workclass	workclass = Private	State-gov	Never-worked	ordinal
Private	1	0	0	1
State-gov	0	1	0	2
Never-worked	0	0	1	3

In ordinal encoding, each unique category is easily reversible, and assigned an integer value.
In binary (or one-hot encoding), no such ordinal relationship exists. This assumption of natural ordering can mean there is poor performance or unexpected results.

Euclidean distance used for "real-valued feature vectors"

$$D(u,v)^2 = \|u-v\|^2 = (u-v)^T(u-v) = \sum_{i=1}^{d}(u_i - v_i)^2$$

The Euclidean distance with ordinal values will have a smaller immediate value. Using binary as is creates a larger distance to sum over.

data vector

Q5. After counting the income values and programming the steps:

1. print the values within income that have the value [1]
   1967
2. print the values within income that have the value [0]
   6033
3. Above 50k Income
   1967
4. Not above 50k income is just the difference
   6033
5. Total Count = above 50k + Not Above 50k
6. Total Count
   8000
7. percent Above 50k = (Above 50k / Total Count) * 100
   percent Above 50k
   24.587000 %

datarobot.com

→ How much data is needed to train a good model?
It depends.

Q6. Norms and Distances [2pt]

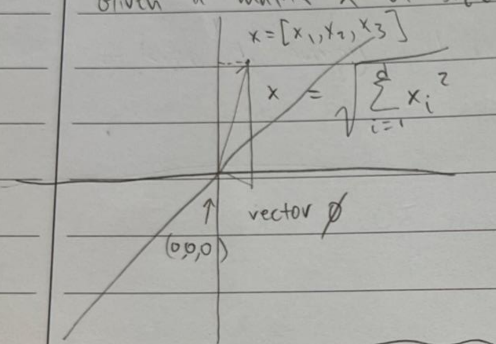
Q6 cont.

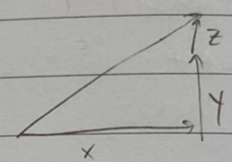
$\|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$ Euclidean norm

A norm is a function that takes a vector as an input and returns a scalar value that can be interpreted as size, length, or magnitude of the vector. → stack overflow.

x z L_2 norm

Given a matrix X of size $n \times d$ and a vector z of size d





$\sqrt{x^2 + y^2 + z^2}$

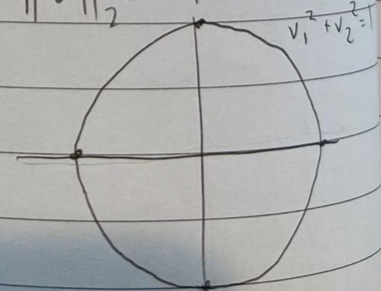
$\|z\|_2 = \sqrt{\sum_{i=1}^{d} z_i^2}$

Vector norm (L_2)

$p=2$ $\sqrt{v_1^2 + \dots + v_n^2}$

$p=1$ $|v_1| + \dots + |v_n|$

$\|v\|_2 = 1$



Q6 cont.

Euclidean distance

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

proof that Euclidean distance $\equiv$ L_2 norm

$$\|x\|_2 - \|z\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2 - \sum_{i=1}^{d} z_i^2}$$

$$= \sqrt{\sum_{i=1}^{d} (x_i - z_i)^2}$$

$$= \sqrt{(x_1 + x_2 + \ldots + x_d)^2 - (z_1 + z_2 + \ldots + z_d)^2}$$

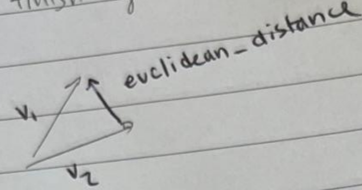
$$\|x - z\|_2 = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2 + \ldots (x_d - z_d)^2}$$

The $(x - z)$ is its own value

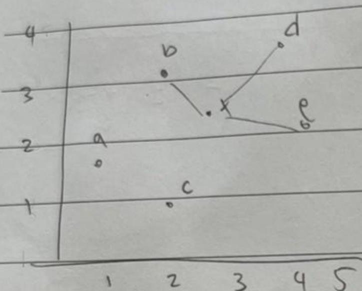
Therefore, the Euclidean distance between x and z can be written as an L_2 norm

Oct 15, 2022

Q7. finish get_nearest neighbors



$$\begin{bmatrix} 0 & -10 \\ \pi & 5 \end{bmatrix} \quad n = 2$$



Data		k	Query
a	(x, y)	1	x
b			
c			
d			
e			
f			
g			

expect:

	Train	sort → d	test row
0	[1, 0, 2]	✓	[1, 4, 2]
1	[3, -2, 4]	✓	
2	[5, -2, 4]	✓	
3	[4, 2, 1.5]	✓	
4	[3.2, π, 2]	✓	
5	[1.5, 0, 1]		

idx changes!

Part 2 questions:

In knn.py (within a .zip file)

Part 3 questions:
1. 10 hours
2. moderate to difficult
3. mostly alone (math) with Discord help and discussion (code)
4. 70%
5. No