Umair Afzal
HW1
CS589

Collaborators:
Debankita,
Suraj

1. Probability and Estimation [30 points]

1. .

P&E

1. $P(y=1|\vec{w}) = \frac{1}{z} P(y=1) P(\vec{w}|y=1)$ $\quad \vec{x} = (0,1,1)$

$= \frac{1}{z}(0.3)(0.3)(0.5)$

$= \frac{1}{z}(0.045)$

$P(y=0|\vec{w}) = \frac{1}{z} P(y=0) P(\vec{w}|y=0)$

$= \frac{1}{z}(0.7)(0.5)(0.1)$

$= \frac{1}{z}(0.035)$ $\quad P(y=0|\vec{w})$ has higher posterior probability

$P(y=1|\vec{w}) + P(y=0|\vec{w}) = z$

$z = 0.045 + 0.035 = \frac{8}{100}$

$P(y=1|\vec{w}) = \frac{0.045}{8/100} = 0.5625$

The naive assumption in this problem is that we are incorporating features of the class which are independent from each other.

2. .

P&E

2. Given that we have N samples $x_1 \dots x_N$, variance σ^2 mean μ

We begin by the likelihood $P(x_1 \dots x_N | \mu) =$

$\prod_{i=1}^{N} P(x_i|\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$\log(P(x_1 \dots x_N|\mu)) = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}$

Differintiate with respect to mean

$= \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2}$

$\frac{d}{d\mu}\left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$

$0 - \frac{\cancel{2}(x_i-\mu)}{\cancel{2}\sigma^2}$

$0 = \sum_{i=1}^{N} \frac{(x_i-\mu)}{\sigma^2} = \sum_{i=1}^{N}(x_i-\mu)$

$\sum_{i=1}^{N}\mu = \sum_{i=1}^{N} x_i \longrightarrow N\mu = \sum_{i=1}^{N} x_i$

$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

3. .

P & ~~E~~

3. $\theta_{MLE} = \arg\max_{\theta} P(data|\theta)$

$= \arg\max_{\theta} \log P(data|\theta)$

$\theta_{MAP} = \arg\max_{\theta} P(\theta|data)$

$= \arg\max_{\theta} \frac{P(data|\theta)P(\theta)}{P(data)}$

$= \arg\max_{\theta} P(data|\theta)P(\theta)$

$= \arg\max_{\theta} \log P(data|\theta) + \log P(\theta)$

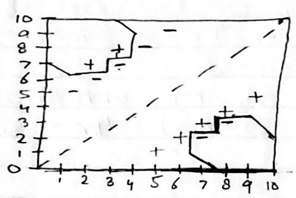
The situations under which MLE & MAP estimators are identical when $P(\theta)$ is 0. Also, they are identical when θ is constant. This happens in the uniform prior when prior probability is a constant function.

2. K Nearest Neighbors [25 points]

1. .

KNN

1)



A very large k will be bad because it will have low variance and high bias. A very small k will be bad because it will have low bias and high variance.

Regarding Bias-variance tradeoff, if the number of nearest neighbours increase, we will have high variance and low bias.

The value of k that minimizes leave-one-out cross-validation error for this dataset is 5. The resulting error is 2/7

2.

Neighbors	Total Score	Time Taken (ms)
3	0.7661954725776303	1000458.3148
5	0.7391429793997188	1025328.7596999999
10	0.6960227289254227	1061917.4712
20	0.6358901533249359	1015361.0788999999
25	0.6119830476889301	1046122.5814

3. .

KNN

3) It is reasonable to use F1-score because it incorporates false positives and negatives. Also, after we calculate the 'recall' and 'precision' we calculate the weighted average. Since the class distribution of our problem is not even, using F1-score as evalution metric will detet weird behaviour and in our problem domain. (unauthorized transactions)

3. Decision Trees [25 points]

1. .

Decision Tree

1. Basically, the criteria used to select a variable for a node when training a decision tree is linked to the information gain that is optimized.
We should not use optimal odering search trough the whole tree as it will take so much time and is possibly not going to work.

2. Code (Not working)
3. NOT ATTEMPTED

4. Model Selection & Hyperparameter Tuning [10 points]

1. .

1. Logistic regression is considered high bias because we try to perform a linear decision boundary and it is possible that it might not exist as a case in the data.
Decision Tree is ——— considered low bias because they maximally overfit to the training data for decision
KNN is considered low bias because we fit the model only to the one nearest point.

2. .

2. Time complexity = $O((k-1)\frac{N}{k})$

When k=5 and k=N, runtime is O(N). However, k=N has a slight greater runtime (means takes a bit longer). It is recommend to use on small data because it is unbiased and have greater variance
For the five-fold cross-validation, we can have a balance in bias and variance. It can have less variance but greater bias.

5. Train Your Best Model [10 points]

1.

Metric	Precision	Recall	F1 Score	AUC
Avg. Validation Set	0.8816437826111635	0.7661521148834581	0.8184761099394919	0.8829909173329682
Full Training Set	0.19936102236421724	0.18682634730538922	0.19289026275115922	0.19341267281629784

I selected the Decision Tree Classifier as it is more suitable for the given data set. When depth is 6, I get the maximum AUC for Avg. Validation Set. You can find the metrics in the Submission Folder (best_model_metrics.txt). If we take greater depths, the model will start to overfit

Sources:
https://www.geeksforgeeks.org/decision-tree-implementation-python/
https://www.datacamp.com/community/tutorials/decision-tree-classification-python
https://wiseodd.github.io/techblog/2017/01/01/mle-vs-map/