CPSC 340 Assignment 3 (due Friday, Oct 12 at 11:55pm)

1 Finding Similar Items

1.1 Exploratory data analysis

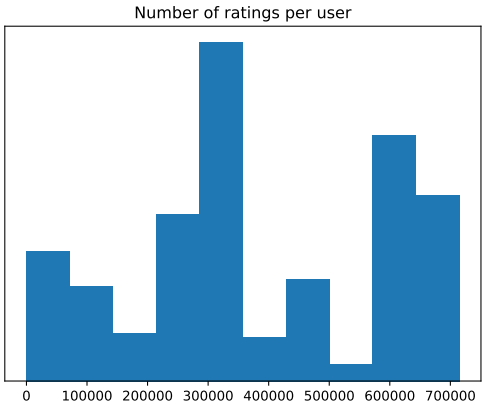
1.1.1 Most popular item

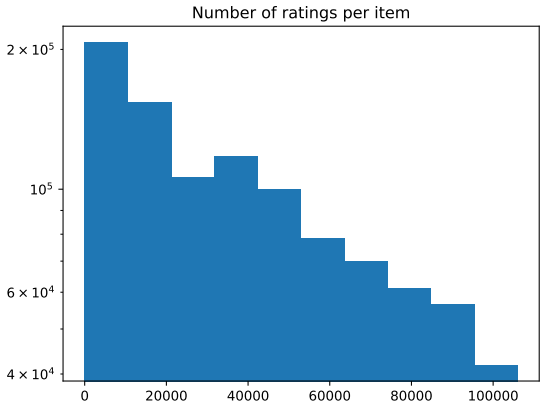
Item: Classic Accessories 73942 Veranda Grill Cover - Durable BBQ Cover with Heavy-Duty Weather Resistant Fabric, X-Large, 70-Inch Number of Stars: 14454

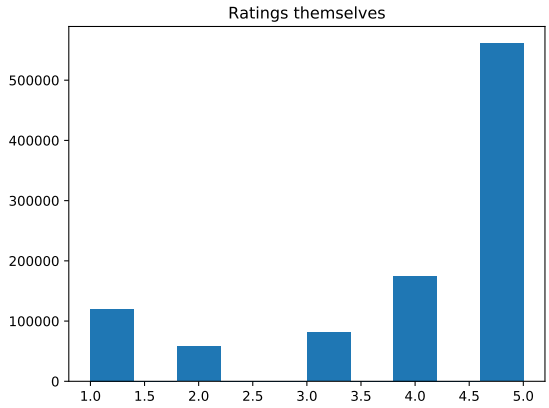
1.1.2 User with most reviews

User ID: A100WO06OQR8BQ, Number of Items: 161

1.1.3 Histograms







2 Matrix Notation and Minimizing Quadratics

2.1 Converting to Matrix/Vector/Norm Notation

1. $||Xw - y||_\infty$
2. $(Xw - y)^T V(Xw - y) + \frac{\lambda}{2}||w||_2$
3. $|Xw - y|_1^2 + \frac{1}{2}A|w|_1$

2.2 Minimizing Quadratic Functions as Linear Systems

1.

$$\begin{aligned}
\frac{1}{2}||w - v||^2 &= \frac{1}{2}(w^T - v^T)(w - v) \\
&= \frac{1}{2}w^T w - w^T v - \frac{1}{2}v^T v \\
&= w - v \\
w - v &= 0 \\
w &= v
\end{aligned}$$

2.

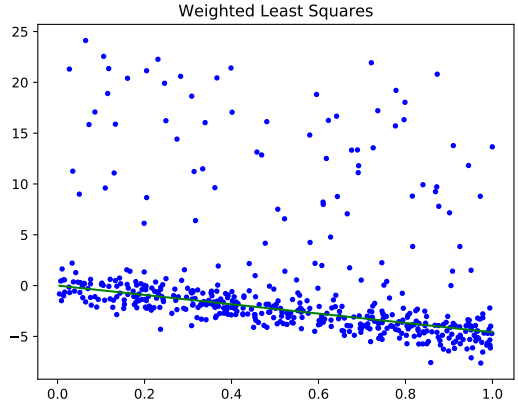
$$\begin{aligned}
\frac{1}{2}||Xw - y||^2 + \frac{1}{2}w^T Aw &= \frac{1}{2}(w^T X^T - y)(Xw - y) + \frac{1}{2}w^T Aw \\
&= \frac{1}{2}w^T X^T Xw - w^T X^T y - \frac{1}{2}y^T y + \frac{1}{2}w^T Aw \\
&= X^T Xw - X^T y + Aw \\
X^T Xw - X^T y + Aw &= 0 \\
w(X^T X + AI) &= +X^T y
\end{aligned}$$

3.

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^{n} v_i(w^T x_i - y_i)^2 + \frac{\lambda}{2}||w - w^0||^2 &= v\frac{1}{2}(w^T X^T - y)(Xw - y) + \frac{\lambda}{2}(w^T - (w^0)^T)(w - w^0) \\
&= v(\frac{1}{2}w^T X^T Xw - w^T X^T y - \frac{1}{2}y^T y) + \frac{\lambda}{2}w^T w - \lambda w^T w^0 + \frac{\lambda}{2}(w^0)^T w^0) \\
&= vX^T Xw - vX^T y - \lambda w - \lambda w^0 \\
vX^T Xw - vX^T y - \lambda w - \lambda w^0 &= 0 \\
vX^T Xw - \lambda w &= vX^T y + \lambda w^0 \\
w(X^T Xw - \lambda I) &= vX^T y + \lambda w^0
\end{aligned}$$

3 Robust Regression and Gradient Descent

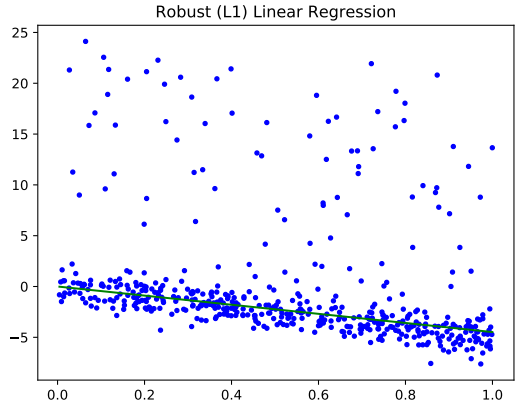
3.1 Weighted Least Squares in One Dimension



3.2 Smooth Approximation to the L1-Norm

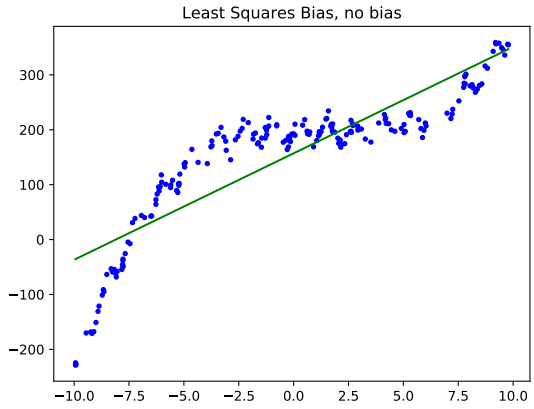
$$
\begin{aligned}
\nabla F &= \frac{d}{dr}[log(exp(r) + exp(-r))] \\
&= \frac{exp(r) - exp(-r)}{exp(r) + exp(-r)} \\
\frac{\partial f}{\partial w_j} &= \sum_{i=1}^{n} x_{ij} \frac{exp(w^T x_i - y_i) - exp(y_i - w^T x_i)}{exp(w^T x_i - y_i) + exp(y_i - w^T x_i)} \\
r_i &= \frac{exp(w^T x_i - y_i) - exp(y_i - w^T x_i)}{exp(w^T x_i - y_i) + exp(y_i - w^T x_i)} \\
&= X^T \frac{exp(Xw - y) - exp(y - Xw)}{exp(Xw - y) + exp(y - Xw)}
\end{aligned}
$$

3.3 Robust Regression



4 Linear Regression and Nonlinear Bases

4.1 Adding a Bias Variable



4.2 Polynomial Basis

p=0 Training error = 15480.5 Test error = 14390.8
p=1 Training error = 3551.3 Test error = 3393.9
p=2 Training error = 2168.0 Test error = 2480.7
p=3 Training error = 252.0 Test error = 242.8
p=4 Training error = 251.5 Test error = 242.1
p=5 Training error = 251.1 Test error = 239.5
p=6 Training error = 248.6 Test error = 246.0
p=7 Training error = 247.0 Test error = 242.9
p=8 Training error = 241.3 Test error = 246.0
p=9 Training error = 235.8 Test error = 259.3
p=10 Training error = 235.1 Test error = 256.3
Generally, as the degree of the polynomial increased, the training error and test error decreased and the amount it decreases becomes less and less as p gets larger.

5 Very-Short Answer Question

1. For a global outlier in K-Means, the outlier will be assigned the nearest mean value and the mean value will be not as close to the other points to accommadate for the global outlier. In clustering, the global outlier will not be assigned to anything as the outlier only consists of itself and there would be no clustering whatsoever.
2. Random restarts are necessary for k-means because the the means in k-means will be initialized randomly. Increasing the number of restarts will decrease the effects of random initialization. It is unnecessary for density clustering because density clustering looks at the points around a radius of given points rather than randomly initialized means.
3. Yes
4. Computing the z score. A problem is that this model assumes that the data is centered around the mean which may not be true.
5. Boxplots. A problem with boxplots is that it looks at only 1 variable at a time

6. Decision Trees. A problem is that we need to label a data set to be of non-outliers and outliers and we may not detect other outliers since it may train on limitied data that does not have all types of outliers.
7. Gradient descent would not be needed as it is easy to use the ordinary linear squares solution since d is small (d=1). Computing the ordinary least squares would only require $O(n)$ time where as $O(nt)$ for gradient descent
8. We only add the column of 1 values to X because we need to accommadate for the y-intercept β. It would not make sense to add this column to the training set for decision trees.
9. If a function is convex, it will have at most 1 stationary/critical point. The critical/stationary point may not exist such as in the case of absolute value function. There will always be a minimizer and a minimum point for convex functions.
10. If the learning rate is too small, we may need many iterations to converge to the best values.
11. If the learning rate is too big, we may skip over the optimal solution.
12. If we were to look for the gradient of an absolute function, it may not exist. The purpose of log-sum-exp is to Smooth out the function so we can compute the gradient for the function.
13. Using a polynomial basis