Maximum Likelihood, Robust Linear Regression

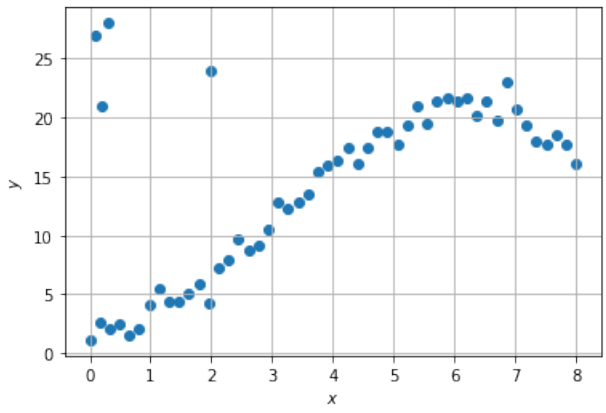
Submit a PDF of your answers to Canvas.

1. A common approach to regression is to imagine the observed data was generated by an additive noise model:

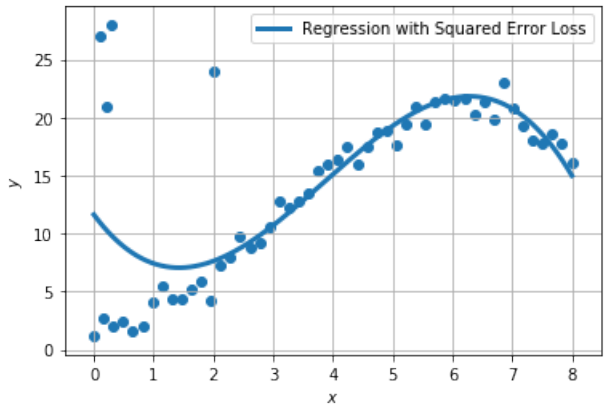
$$y = \boldsymbol{x}^T \boldsymbol{w} + Z$$

where Z is additive noise, $\boldsymbol{x}$ is a feature vector and $\boldsymbol{w}$ is an unknown vector of coefficients. When $Z \sim \mathcal{N}(0, \sigma^2)$, we saw that maximum likelihood estimation for $\boldsymbol{w}$ was equivalent to minimizing the *square error loss*. One shortcoming of a squared error loss function is that it results in a fit that is not robust to outliers, since the loss function is (of course) proportional to the *squared* distance between the true value and the prediction.

The scatter plot below shows a dataset $\mathcal{D} = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ that includes 4 outliers. The outliers are in the upper left of the plot.



Our goal is to find a *robust* degree-3 polynomial that fits the data points. Recall that we can fit a 3-degree polynomial by first creating a feature vector $\boldsymbol{x} = [x^3 \;\; x^2 \;\; x^1 \;\; 1]^T$ for each data point, stacking these as the rows of a $n \times 4$ matrix $\boldsymbol{X}$, and computing $\boldsymbol{w}_{LS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$ (which is the closed form for minimizing the squared error loss function). The resulting curve is shown on the plot below. Note the impact of the outliers.



In this problem, to make our fit more robust to outliers, we will assume that the additive noise follows a Laplace distribution. More precisely, we will assume the data was generated by

$$y = \boldsymbol{w}^T \boldsymbol{x} + Z$$

where $Z \sim \text{Lap}(0, 1)$. Recall the standard Laplace distribution $\text{Lap}(0, 1)$:

$$p(z) = \frac{1}{2} e^{-|z|}$$

a) Find an expression for the distribution of y for a fixed $\boldsymbol{x}$, $\boldsymbol{w}$, denoted $p_{\boldsymbol{w}}(y)$.

b) You collect independent measurements and store them in a dataset $\mathcal{D} = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. Find an expression for the distribution of $y_1, y_2, \ldots$ for a fixed $\boldsymbol{w}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, denoted $p_{\boldsymbol{w}}(\boldsymbol{y})$.

c) Recall that the maximum likelihood estimate of $\boldsymbol{w}$ given $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ is given by

$$\arg\max_{\boldsymbol{w}} p_{\boldsymbol{w}}(\boldsymbol{y})$$

which is equivalent to

$$\arg\max_{\boldsymbol{w}} \log(p_{\boldsymbol{w}}(\boldsymbol{y})).$$

Use your result for $p_{\boldsymbol{w}}(\boldsymbol{y})$ to simplify the expression above as much as possible. Make use of the notation for the ℓ_1 norm: $\sum |a_i| = ||\boldsymbol{a}||_1$.

d) Why is the additive Laplacian noise model more robust to outliers?

SOLUTION:

a) Adding a constant to a pdf shifts the pdf, and $p_{\boldsymbol{w}}(y) = \frac{1}{2}e^{-|y-\boldsymbol{w}^T\boldsymbol{x}|}$.

b) Since the measurements are independent, $p_{\boldsymbol{w}}(\boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{2}e^{-|y_i-\boldsymbol{w}^T\boldsymbol{x}_i|}$.

c) Note that

$$\begin{aligned}
\arg\max_{\boldsymbol{w}} \log p_{\boldsymbol{w}}(\boldsymbol{y}) &= \arg\max_{\boldsymbol{w}} \sum -\log(2) + \sum -|y - \boldsymbol{w}^T\boldsymbol{x}_i| \\
&= \arg\min_{\boldsymbol{w}} \sum |y_i - \boldsymbol{w}^T\boldsymbol{x}_i| \\
&= \arg\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_1.
\end{aligned}$$

The results of the additive Gaussian model and the Laplace model are shown below. Note that unlike the least squares estimate, $\boldsymbol{w}_{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, which has a closed form, we have to use numerical optimization to compute the maximum likelihood estimate of $\boldsymbol{w}$ under the Laplacian noise assumption.