DS203: Programming in Data Science

Regression

Manjesh K. Hanawal

14th Oct 2020

Plain

Previous Lecture:

- Classification vs Regression
- Linear Regression
- Simple and Multiple Linear Regression
- Goodness of fit and Correlation coefficient

This Lecture:

- Model Validation
- Regression Dignostics
- Outlier detection
- Loss functions for Robustness
- Ridge and LASSO Regression
- Logistic Regression

Solution of Multiple Linear Regression

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{m} (y_i - x_i \beta^T)^2$$

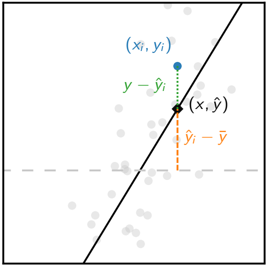
$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Least Squared Estimator (LSE)

For any $x \in \mathcal{X}$, the predicted value is $\hat{y} = \hat{\beta} x$

Model Evaluation:

Suppose every point y_i is very close to $\bar{y} \implies y_i$ does not dependent much on x_i and there is not much random error.



$$y_i - \bar{y} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{explained by model}} + \underbrace{(y_i - \hat{y}_i)}_{\text{not explained by model}}$$

Coefficient Determination

$$\underbrace{\sum_i (y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{\text{SSM}} + \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{\text{SSE}}$$

$$1 = \underbrace{\frac{SSM}{SST}}_{r^2} + \underbrace{\frac{SSE}{SST}}_{1-r^2}$$

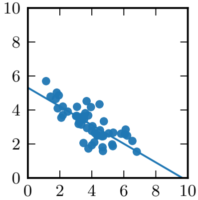
- r^2 is called the coefficient of determination (square of coefficient of correlation!)
- Captures the fraction of variability explained by model
- It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph
- closer to 1, the better.

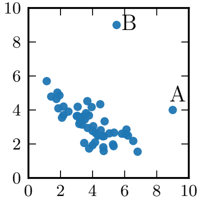
Regression Dignostics

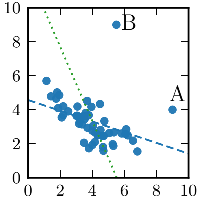
- Real dataset can have some data points that are too noisy. How to handle noisy data points?
- Is minimizing the Mean Squared Error always best?
- What if we have more features than data points?
- In real datasets some feature may be similar to each other (correlated). Should all the features be given importance?

Handling outliers

- Outlier is any point that is ‘far away’ from the rest of the data
- Leverage of a data point is the quantitative description of how far it is from the rest of the points in the x-axis
- Influential point is an outlier with high leverage that significantly affects the slope.







Leverage and influential points

Presence of Influential points Indicate

- Need for more data aggregation: If most of the data is concentrated in some region, may be sampling method is flawed we might have missed some data points.
- Points are noisy : If there is no flaw in data collection techniques, then they must be removed.

Definition of leverage points:

- $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$. X is data matrix
- Hat matrix: $H = X(X^T X)^{-1} X^T$, leverage of points x_i is H_{ii}, i.e., ith diagonal element. In one dimension

$$H_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$$

- Relatively large values of x_i's have larger leverage.

Quantifying Influence of a point: Cook's distance

Algorithm:

- Take a point and fit the model with and without it.
- More different they are, more influential is the model.
- Influence of point i is then given by Cook's distance

$$D_i = \frac{\frac{1}{d}\sum_j (\hat{y}_j - \hat{y}_{j(-i)})^2}{\frac{1}{m}\sum_j (y_j - \hat{y}_j)^2}$$

- Remove all points for which $D_i \geq \alpha$ (threshold)
- Computation of D_i needs retraining. Can be avoided!

$$D_i = \frac{1}{d}\frac{1}{MSE}\frac{H_{ii}}{(1 - H_{ii})^2}(y_i - \hat{y}_i)^2$$

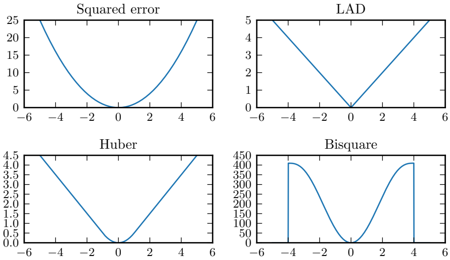
Robustness in Optimization

- Standard linear regression (based on LSE minimization) is affected by outliers
- Removing outliers can be cumbersome.
- How to make linear regression more robust?
- Tweak the objective of optimization?

- Define $r_i = y_i - x_i\beta^T$ as the error at point
- L_β be a function defined on r_i
- Total loss is $\sum_j L_\beta(r_j)$. Minimize $\sum_j L_\beta(r_j)$ over β.

$L_\beta(r) = r^2$ gives the LSE function. Gives more penalty to large errors (outliers). Example: $L_\beta(5) = 25$ and $L_\beta(10) = 100$.
Tries hard to fit the outliers!

Loss functions



Least Absolute Deviations (LAD): $L_\beta(r) = |r|$.

- Avoids excessively penalizing outliers
- Not differentiable at origin ($r = 0$). Hard to optimize!

Huber Loss:

$$L_\beta(r) = \begin{cases} r^2/2 & |r| < k \\ k(|r| - k/2) & |r| \geq k. \end{cases}$$

- Similar to LAD but avoids sharp change at origin making it differentiable everywhere. Helps in optimization

Distribution view of Robustness

- The optimization problem could be is governed by the type of randomness (noise)
- When noise is Gaussian distributed, the best optimizer is obtained by minimizing the Squared Error Loss (Recall MLE!)
- When noise is Laplacian, the best optimizer is obtained by minimizing the LAD (verify!)

Sparsity in Optimization Solution:

- There could be more features than data points
- Features could be similar to each other or dependent
- We want the β to be sparse, i.e., $\beta_k = 0$ for many k

Ridge Regression:

$$\min_{\beta} \left[\underbrace{\sum_{i=1}^{m} (y_i - x_i\beta^T)^2 +}_{\text{data term}} \quad \underbrace{\lambda \sum_{k=1}^{d} \beta_k^2}_{\text{regularization term}} \right]$$

- λ is a non-negative parameter trade off error and the how small we want coefficients to be!

Solution of Ridge Regression

$$\min_{\beta}\left[\sum_{i=1}^{m}(y_i - x_i\beta^T)^2 + \lambda\sum_{k=1}^{d}\beta_k^2\right]$$

$$\hat{\beta} = X(X^TX - \lambda I)X^T$$

(X^T*X + lambda*I)^(-1)*X^T*Y

- If $\lambda = 0$, we get solution of multiple linear regression
- If λ is very large $\hat{\beta} = 0$.
- $\hat{\beta}$ is not always sparse!

Contrived Example:

- Assume all feature are same
- $\beta = (1, 1, 1)$ and $\beta = (0, 0, 3)$ gives same label to the sample
- $\beta = (1, 1, 1)$ has less regularization penalty than $\beta = (0, 0, 3)$!

LASSO Regression

$$\min_{\beta}\left[\underbrace{\sum_{i=1}^{m}(y_i - x_i\beta^T)^2 +}_{\text{data term}} \quad \underbrace{\lambda\sum_{k=1}^{d}\mathbb{1}_{\{\beta_k^2\neq 0\}}}_{\#\text{ of nonzero terms}}\right]$$

- Imposes penalty for nonzero values of β
- The optimization problem is intractable
- No efficient method is known for computing
- Instead, solve the following:

$$\min_{\beta}\left[\sum_{i=1}^{m}(y_i - x_i\beta^T)^2 + \lambda\sum_{k=1}^{d}|\beta_k|\right]$$

l_1 norm minimization or Lease Absolute Shrinkage and Selection Operator (LASSO)

Bayesian Connection

- We treated β is fixed (unknown) parameter: $(y = X\beta^T + \epsilon)$
- What if we treat it as a random quantity? (Bayesian view!)
- Suppose we have some prior belief about β before (X, y) is known, we encode its as **prior distribution** $P(\beta)$.
- Once (X, y) is observed, we can compute more informed distribution on β, call **posterior distribution**
- By Bayes rule $P(\beta|(X,y))P((X,y)) = P(\beta)P((X,y)|\beta)$, or $P(\beta|(X,y)) \propto P(\beta)P((X,y)|\beta)$
- If $P(\beta) \sim \mathcal{N}(0, 1/2\lambda^2)$, we get Ridge regression!
- If $P(\beta) \sim exp\{-\lambda|\beta|\}$, we get LASSO regression!

Linear Regression for classification

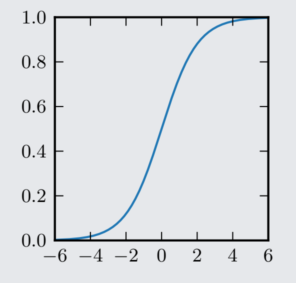
- Focus on binary classification $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{1, -1\}$
- For some given threshold α

$$y = \begin{cases} 1 & \text{if} \quad x\beta^T > 0 \\ 0 & \text{otherwise.} If \end{cases}$$

- $P(Y = 1|X = x) = g^{-1}(x\beta^T)$, where $g(\cdot)$ is the link function.
- **Logistic function**

$$g^{-1}(\eta) = \frac{1}{1 + e^{-u}}.$$

Sigmoid function



- **Logit: (inverse of sigmoid)**

$$\log \frac{P(Y=1|X=x)}{1-P(Y=1|X=x)} = x\beta^T$$

Solving Logistic Regression

$$\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$$

$$L_\beta(\mathcal{S}) = \prod_{i=1}^{m} \left(\frac{1}{1 + e^{-x_i \beta^T}}\right)^{y_i} \left(1 - \frac{1}{1 + e^{-x_i \beta^T}}\right)^{1-y_i}$$

Taking Log both sides:

$$log(L_\beta(\mathcal{S})) = \sum_{i=1}^{m} \left[y_i \log\left(\frac{1}{1 + e^{-x_i \beta^T}}\right) + (1 - y_i) \log\left(\frac{e^{-x_i \beta^T}}{1 + e^{-x_i \beta^T}}\right)\right]$$