COMP 562 – Lecture 5

Feature Scaling -- Feature Scaling

- Idea: gradient ascent/descentalgorithm tends to work better if the features are on the **same scale**

When contours are skewed then learning steps would take longer to converge due to oscillatory behaviour

Feature Scaling -- Centering

Center features by removing the mean

$$\mu_i = \frac{1}{N}\sum_{k=1}^{N} x_{i,k}$$

$$x_{i,j} = x_{i,j} - \mu_i$$

This makes each feature's mean equal to 0. Compute the mean first, then subtract it!

Feature Scaling -- Standardizing

Standardize centered features by dividing by the standard deviation

$$\sigma_i = \sqrt{\frac{1}{N-1}\sum_j x_{i,j}^2}$$

$$x_{i,j} = \frac{x_{i,j}}{\sigma_i}$$

Note that standardized features are first centered and then divided by their standard deviation

Transform your data to a distribution that has a mean of 0 and a standard deviation of 1 (z-score)

Feature Scaling -- Normalizing

Alternatively, **normalize** centered features by dividing by their norm

$$r_i = \sqrt{\sum_j x_{i,j}^2}$$

$$x_{i,j} = \frac{x_{i,j}}{r_i}$$

Note that normalized features are first centered and then divided by their norm

Normalization transforms your data into a range between 0 and 1 regardless of the data set size

Feature Scaling Benefits

1. Centering
   A. β_0 is equal to the mean of the target variable
   B. Feature weights β now tell us how much does feature's departure from mean affect the target variable
2. Standardization
   A. All the features are on the same scale and their effects comparable
   B. Interpretation is easier: βs tell us how much departure by single standard deviation affects the target variable
3. Normalization
   A. Scale of features is the same, regardles of the size of the dataset
   B. Hence weights learend on different sized datasets can be compared
   C. However, their combination might be problematic -- certainly we don't trust weights learned on few samples

Classification -- Bernoulli View

We can model a target variable $y \in \{0, 1\}$ using Bernouli distribution

$$p(y = 1|\theta) = \theta$$

We note that θ has to be in range $[0, 1]$

We cannot directly take weighted combination of features to obtain θ

We need a way to map $\mathbf{x}^T\beta \in \mathbb{R}$ to range $[0, 1]$

Some Useful Equalities Involving Sigmoid

Definition:

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Recognize the alternative way to write it:

$$\sigma(z) = \frac{\exp z}{1 + \exp z}$$

Complement is just flip of the sign in the argument

$$\sigma(-z) = 1 - \sigma(z)$$

Log ratio of probability (log odds)

$$\log \frac{\sigma(z)}{\sigma(-z)} = z$$

Using Sigmoid to Parameterize Bernoulli

$$p(y = 1|\theta) = \theta$$

Sigmoid "squashes" the whole real line into range $[0, 1]$

Hence we can map weighted features into a parameter θ

$$\theta = \sigma(\beta_0 + \mathbf{x}^T \beta)$$

and use that θ in our Bernoulli

$$p(y = 1|\theta = \sigma(\beta_0 + \mathbf{x}^T \beta)) = \sigma(\beta_0 + \mathbf{x}^T \beta)$$

Logistic Regression -- Binary Classification

In logistic regression we model a binary variable $y \in \{-1, +1\}$

$$p(y=+1|\mathbf{x}, \beta_0, \beta) = \sigma\left(+(\beta_0 + \mathbf{x}^T\beta)\right)$$
$$p(y=-1|\mathbf{x}, \beta_0, \beta) = 1 - \sigma\left(-(\beta_0 + \mathbf{x}^T\beta)\right) = \sigma\left(-(\beta_0 + \mathbf{x}^T\beta)\right)$$

This is equivalent to

$$p(y|\mathbf{x}, \beta_0, \beta) = \sigma\left(y(\beta_0 + \mathbf{x}^T\beta)\right) = \frac{1}{1 + \exp\left\{-y(\beta_0 + \mathbf{x}^T\beta)\right\}}$$

Q: Does above formula work for $y \in \{0, 1\}$?

Logistic Regression -- Decision Boundary

$$p(y=1|\mathbf{x}, \beta_0, \beta) = \sigma\left((\beta_0 + \mathbf{x}^T\beta)\right) = \frac{1}{1 + \exp\left\{-(\beta_0 + \mathbf{x}^T\beta)\right\}}$$

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

- Suppose predict "$y = 1$" if

$$p(y=1|\mathbf{x}, \beta_0, \beta) \geq 0.5 \rightarrow \beta_0 + \mathbf{x}^T\beta \geq 0$$

- Then predict "$y = -1$" if

$$p(y=1|\mathbf{x}, \beta_0, \beta) < 0.5 \rightarrow \beta_0 + \mathbf{x}^T\beta < 0$$

- Hence, the decision boundary is given by $\beta_0 + \mathbf{x}^T\beta = 0$

Q: What does this decision boundary equation describe?

Logistic Regression -- Log-Likelihood

Probability of a single sample is:

$$p(y|\mathbf{x}, \beta_0, \beta) = \frac{1}{1 + \exp\left\{-y(\beta_0 + \mathbf{x}^T \beta)\right\}}$$

Likelihood function is:

$$\mathcal{L}(\beta_0, \beta|\mathbf{y}, \mathbf{x}) = \prod_i \frac{1}{1 + \exp\left\{-y_i(\beta_0 + \mathbf{x}_i^T \beta)\right\}}$$

Log-likelihood function is:

$$\log \mathcal{L}(\beta_0, \beta|\mathbf{y}, \mathbf{x}) = -\sum_i \log\left\{1 + \exp\left\{-y_i(\beta_0 + \mathbf{x}_i^T \beta)\right\}\right\}$$

Follow the same recipe as before to find βs that maximize the Log-likelihood function