Lecture 6: Regression
Modeling Social Data, Spring 2017
Columbia University

Tejas Dharamsi

February 25, 2017

1 Introduction

1.1 What is Regression?

- best-fit line
- best-fit curve
- finding patterns in the data
- function we learn

1.2 Formal Definition:

"...to understand as far as possible with the available data how the conditional distribution of the response y varies across subpopulations determined by the possible values of the predictor or predictors."

1.3 Purpose of Regression

- Describe - compact summary of outcomes
- Predict - future of unobserved conditions
- Explain - associations, interpret effects of coefficients.

1.4 Goal:

Flexible models to describe what we have seen before simple enough to generalize future outcome.

1.5 Framework:

- Specify the outcome + predictors, and form of the model relating them
- Define a loss fn. that quantifies how close a model's prediction are to observed outcomes.
- Develop an algorithm to fit the model to the observations by minimizing this loss.
- Assess model performance + interpret results.

2 Math

Y is a vector of outcome.
X is a matrix of input, each column corresponds to a feature (age,gender) as shown in Fig. 1
Loss function : how well function f predicts the outputs from inputs
$L_i[f] = (y - \hat{y})^2$

$$L[f] = 1/N * \sum_{i=1}^{n} L_i[f] \tag{1}$$

$$L[f] = 1/N * \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

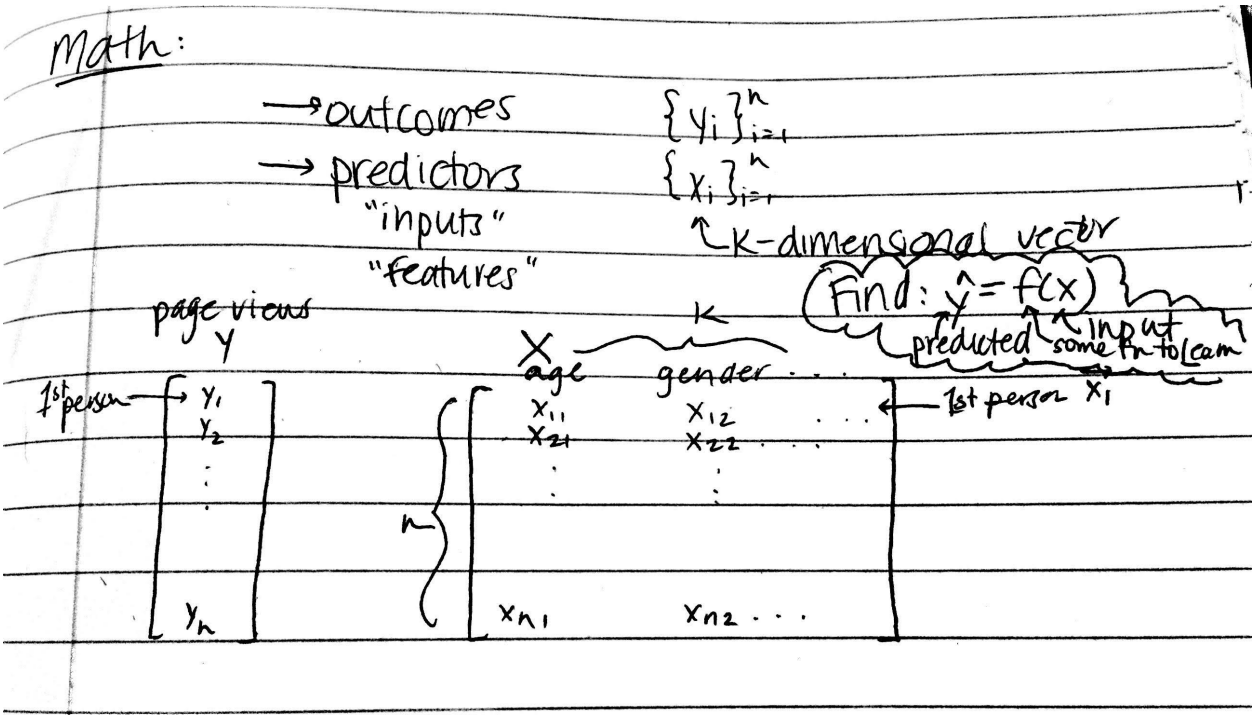

Figure 1: Representing the data for Section Math

2.1 Maximum Likelihood

:

Assume some family of probabilistic model generated the data. Find the model under which the observed data are most likely.

$f^* = \text{argmax}_f P(D|f)$

Common assumption : $y = f(x_i) + E_i$

$P(E_i|f) = P(y_i - f(x_i)|f)$

$= (1 / \sqrt{2\pi\sigma^2}) * \exp^{(-1/2\sigma^2)*(y_i - f(x_i))^2}$

$$P(D|f) = \prod_{i=1}^{n} P(D_i|f) \tag{3}$$

$$P(D|f) = \prod_{i=1}^{n} (1/\sqrt{2\pi\sigma^2}) * exp^{(-1/2\sigma^2)*(y_i - f(x_i))^2} \tag{4}$$

Simplifying Likelihood Fn.

$$P(D|f) = (2\pi\sigma^2)^{-n/2} * exp(-1/2\sigma^2 * \sum_{i=1}^{n} (y_i - f(x_i))^2) \tag{5}$$

$$LogP(D|f) = -n/2 * (2\pi\sigma^2) - 1/2\sigma^2 * \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{6}$$

Maximizing squared loss = maximising (log) likelihood, assuming additive gaussian noise. (errors are normally distributed)

Choice for f: $\hat{y} = f(x) = w.X = w_1x_1 + w_2x_2 + ... + w_nx_n$

$$w^* = argmin_w \sum_{i=1}^{n} (y_i - w.x_i) \tag{7}$$

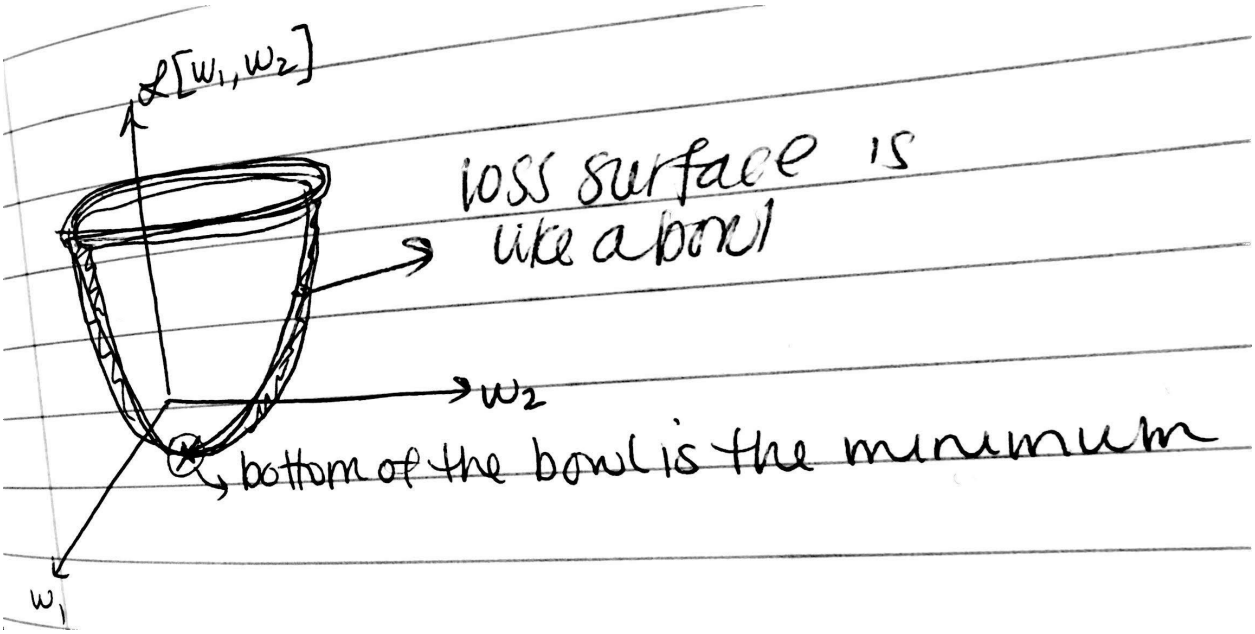

Figure 2: Loss Surface

Loss Surface is easy to visualise in 2D, difficult with high dimensions. One shouldnt do a manual search for the bottom of the loss surface

$$L[f] = 1/N * \sum_{i=1}^{n} (y_i - w.x_i)^2 \tag{8}$$

Differentiating -(8) wrt w and equating with 0.

$$0 = -2/N * \sum_{i=1}^{n} (y_i - w.x_i)x_i \tag{9}$$

Equation -(9) can be written in vectorized notation as :

$$0 = X^T * (y - Xw) \tag{10}$$

Simplifying:

$\hat{w} = (X^TX)^{-1}X^Ty$

$\hat{w}$ is the best estimate of w. Time complexity to find $\hat{w}$ is $O(NK^2 + K^3)$

If the data is very wide, this would be very time consuming.

2.2 Gradient Descent

WRT to Loss Surface

1. Start somewhere random
2. Follow the negative gradient to the bottom of the surface.

w = w - eta * (gradient)
gradient is same as equation - (9)

eta is the step size: large = big steps, small = slow progression
Computational Costs:
Xw = O(KN)
Y-Xw=O(N)
X^T(Y-Xw) =O(KN)
Time Complexity is O(KN + N + KN) = O(KN)

Gradient Descent would be better for higher dimension.

2.3 Stochastic Gradient Descent

Same intuition (follow the gradient) but follow the approximate gradient.

$$gradientofLossfunction(dL/dw) = -2/N * \sum_{i=1}^{n} (y_i - w.x_i)x_i \tag{11}$$

Take a random sample n very very small as compared to N examples to estimate the gradient.

$$gradientofLossfunction(dL/dw) = -2/n * \sum_{i=1}^{n} (y_i - w.x_i)x_i \tag{12}$$

Time Complexity is O(KN + N + KN) = O(KN)

On avg you are correct + close to the bottom. Choosing eta becomes tricky, make it smaller as you go closer to convergence

2.4 2nd order derivative

In methods like newton method, second order derivative of loss function is calculated to determine the curvature of the surface. More computation to take 2nd order derivative but saves time to reach to minimum. Faster steps on flatter surface and smaller steps on steep surface.

2.5 Some Additional Points

1. In highly skewed data, used median over mean.
2. Median and Geometric Mean in log scale are pretty close.
3. Log scale is preferred while working with heavy tail distribution.