

DSC 40A

Lecture 10

Least Squares Regression, pt. I

Last Time

- ▶ How do we make predictions using multiple features?
- ▶ Assume a linear prediction rule:

$$\begin{aligned}H(x_1, \dots, x_d) &= w_0 + w_1 x_1 + w_2 x_2 + \dots + w_d x_d \\&= \text{Aug}(\vec{x}) \cdot \vec{w}\end{aligned}$$

- ▶ We found the normal equations:

$$X^T X \vec{w} = X^T \vec{y}$$

- ▶ Solving the normal equations for \vec{w} gives the best-fitting prediction rule.

Today

- ▶ Interpreting the results.
- ▶ How do we fit prediction rules like $H(x) = w_2x^2 + w_1x + w_0$?
- ▶ Least squares **classification**.

Interpreting \vec{w}

- ▶ With d features, \vec{w} has $d + 1$ entries.
- ▶ w_0 is the **bias**.
- ▶ w_1, \dots, w_d each give the **weight** of a feature.

$$H(\vec{X}) = w_0 + w_1 x_1 + \dots + w_d x_d$$

- ▶ Sign of w_i tells us about relationship between i th feature and outcome.

Example: Predicting Sales

- ▶ For each of 26 stores, we have:
 - ▶ net sales,
 - ▶ size (sq ft),
 - ▶ inventory,
 - ▶ advertising expenditure,
 - ▶ district size,
 - ▶ number of competing stores.
- ▶ Goal: predict net sales given size, inventory, etc.
- ▶ To begin:

$$H(\text{size, competitors}) = w_0 + w_1 \times \text{size} + w_2 \times \text{competitors}$$

Discussion Question

What will be the sign of w_1 and w_2 ?

- A) $w_1 = +$, $w_2 = -$
- B) $w_1 = +$, $w_2 = +$
- C) $w_1 = -$, $w_2 = -$
- D) $w_1 = -$, $w_2 = +$

$$H(\text{size}, \text{competitors}) = w_0 + w_1 \times \text{size} + w_2 \times \text{competitors}$$

(DEMO)

Discussion Question

Which has the greatest effect on the outcome?

- A) size: $w_1 = 16.20$
- B) inventory: $w_2 = 0.17$
- C) advertising: $w_3 = 11.53$
- D) district size: $w_4 = 13.58$
- E) competing stores: $w_5 = -5.31$

Which features are most “important”?

- ▶ **Not necessarily** the feature with largest weight.
- ▶ Features are measured in different units, scales.
- ▶ We should **standardize** each feature.

Standard Units

- ▶ To standardize (z-score) a feature, subtract mean, divide by standard deviation.
- ▶ Example: 10, 20, -30, 5, 15
 - ▶ Mean: 4
 - ▶ Standard Dev: $\sqrt{\frac{1}{5} \sum (x_i - \bar{x})^2} \approx 17.7$
 - ▶ Standardized:

$$\frac{10 - 4}{17.7} = 0.34, \quad \frac{20 - 4}{17.7} = 0.90, \quad \frac{-30 - 4}{17.7} = -1.92,$$

$$\frac{5 - 4}{17.7} = 0.06, \quad \frac{15 - 4}{17.7} = 0.62$$

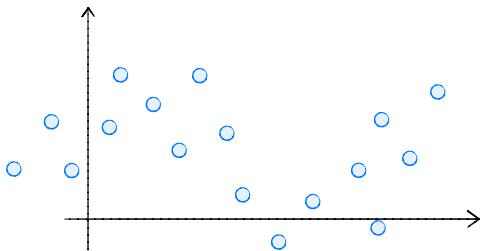
Standard Units

- ▶ Standardize each feature (store size, inventory, etc.) separately.
- ▶ No need to standardize outcome (net sales).
- ▶ Solve normal equations. The resulting w_0, w_1, \dots, w_d are called the **standardized regression coefficients**.
- ▶ They can be directly compared to one another.

(DEMO)

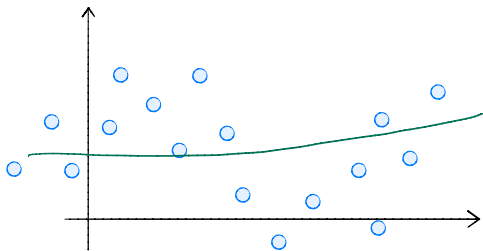
Fitting Non-Linear Patterns

- Fit a 4th-order polynomial to the data:



- We know how to fit rules of the form $H(x) = w_1 x^4 + w_0$.
 - Define $z_i = x_i^4$.
 - Use $w_1 = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})^2}$ and $w_0 = \bar{y} - w_1 \bar{z}$.

The Result



- ▶ The rule $H(x) = w_1x^4 + w_0$ **underfits** the data.
- ▶ We need a more complicated rule:

$$H(x) = w_4x^4 + w_3x^3 + w_2x^2 + w_1x + w_0$$

The Trick

- ▶ Treat x , x^2 , x^3 , x^4 as different features.
- ▶ Create design matrix:

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & x_n^4 \end{pmatrix}$$

- ▶ Solve $X^T X \vec{w} = X^T \vec{w}$ for \vec{w} , as usual.
- ▶ Works for more than just polynomials.

(DEMO)

Polynomial Regression

- ▶ More complicated patterns can be fit with higher-order polynomials.
- ▶ If there are n points, a $n + 1$ degree polynomial can fit them exactly.
- ▶ But for high-order polynomials, it becomes **very hard** to solve the normal equations (numerical accuracy).

Polynomial Regression with Multiple Features

- Suppose we want to fit a rule of the form:

$$\begin{aligned}H(\text{size}, \text{competitors}) &= w_0 + w_1 \text{size} + w_2 \text{size}^2 \\&\quad + w_3 \text{competitors} + w_4 \text{competitors}^2 \\&= w_0 + w_1 s + w_2 s^2 + w_3 c + w_4 c^2\end{aligned}$$

- Make design matrix:

$$X = \begin{pmatrix} 1 & s_1 & s_1^2 & c_1 & c_1^2 \\ 1 & s_2 & s_2^2 & c_2 & c_2^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & s_n & s_n^2 & c_n & c_n^2 \end{pmatrix}$$

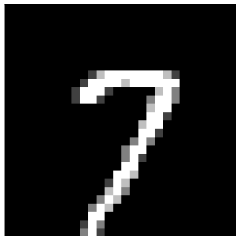
Where c_i and s_i are the competitors and size of the i th store.

Regression vs. Classification

- ▶ **Regression**: predict a number
 - ▶ Examples: salary, store sales, height of child
- ▶ **Classification**: predict a *class*, or *group label*.
 - ▶ is this person at high risk of disease (yes/no)?
 - ▶ what type of tree is in this (pine, elm, oak, etc.)?

Binary Classification

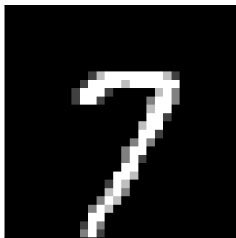
- ▶ There are two possible classes.
- ▶ Example: handwritten digits. Is image a 7, or a 3?



- ▶ Data: images $\vec{x}^{(i)}$, **labels** $y_i = 1$ if a seven, $y_i = 0$ if a three.

Images as Feature Vectors

- ▶ We can pack an image into a feature vector.
- ▶ Each feature is the intensity of a particular pixel.
- ▶ Example: a 28×28 image has 784 pixels, becomes a vector in \mathbb{R}^{784} .



Decision Rule

- ▶ We want a rule $H(\vec{x})$ that takes in images and outputs:
 - ▶ 1 if image is a seven
 - ▶ 0 if image is a three
- ▶ We'll use a linear decision rule:

$$\begin{aligned} H(\text{image}) &= w_0 + w_1 \times (\text{pixel 1}) + \dots + w_{784} \times (\text{pixel 784}) \\ &= \text{Aug}(\vec{x}) \cdot \vec{w} \end{aligned}$$

- ▶ Minimize MSE, same solutions:

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \sum_{i=1}^n \left(\text{Aug}(\vec{x}^{(i)}) \cdot \vec{w} - \vec{y}_i \right)^2 \qquad X^T X \vec{w} = X^T \vec{y}$$

Least Squares Classification

- ▶ Our prediction $H(\vec{x})$ will not be 0 or 1 exactly.
- ▶ If $H(\vec{x}) > \frac{1}{2}$, we'll claim it is a 1; else, a 0.

(DEMO)

Least Squares Classification

- ▶ Square loss is good for regression: want $H(\vec{x})$ close to right answer.
- ▶ Not great for classification.
- ▶ If real class is 1, and $H(x) = 10$, great!
- ▶ If real class is 1, and $H(x) = -1$, not great.
- ▶ Better loss functions: hinge loss, logistic loss, etc.

