Linear regression

COMS 4771 Fall 2023
Dartmouth student dataset
Dataset of 750 Dartmouth students’ (first-year) college GPA

Mean 2.47
Standard deviation 0.75

https://chance.dartmouth.edu/course/Syllabi/Princeton96/ETSVvalidation.html
Dartmouth dataset also has high school GPA of each student

Question: Is high school GPA predictive of college GPA?
Attempting to exploit “local regularity” using NN.

The diagram shows the relationship between high school GPA and college GPA. There are data points for each combination of high school GPA and college GPA. The diagram includes lines indicating the expected college GPA given the high school GPA and predictions from nearest neighbors (NN). The expected value is denoted as $E[\text{college GPA} | \text{high school GPA}]$.

- "1"-NN prediction
- "3"-NN prediction
Possible “global” modeling assumption:

- Increase in high school GPA by $\Delta$ should give an increase in (expected) college GPA by $\propto \Delta$

- In other words,

$$\mathbb{E}[\text{college GPA} \mid \text{high school GPA}]$$

is ______ function of high school GPA
Least squares linear regression
$f : \mathbb{R} \to \mathbb{R}$ is **linear** if it is of the form

$$f(x) = mx + b$$

for some parameters $m, b \in \mathbb{R}$
Problem: given a dataset $S$ from $\mathbb{R} \times \mathbb{R}$, find (parameters of) a linear function $f(x) = mx + b$ of minimal \underline{sum of squared errors (SSE)}

$$
\text{sse}[m, b] = \sum_{(x,y) \in S} (mx + b - y)^2
$$

Method of solution is called \underline{ordinary least squares (OLS)}
Minimizers of SSE must be zeros of the two partial derivative functions:

\[
\frac{\partial \text{sse}}{\partial m}[m, b] = 2 \sum_{(x,y) \in S} (mx + b - y)x = 0
\]

\[
\frac{\partial \text{sse}}{\partial b}[m, b] = 2 \sum_{(x,y) \in S} (mx + b - y) = 0
\]

Two linear equations in two unknowns

Together, the equations are called the normal equations
Equivalent form:

\[
\begin{align*}
\text{avg}(x^2) \, m & \quad + \quad \text{avg}(x) \, b \\
\text{avg}(x) \, m & \quad + \quad b
\end{align*} = \text{avg}(xy)
\]

where

\[
\begin{align*}
\text{avg}(x) &= \frac{1}{|S|} \sum_{(x,y) \in S} x, \\
\text{avg}(x^2) &= \frac{1}{|S|} \sum_{(x,y) \in S} x^2, \\
\text{avg}(xy) &= \frac{1}{|S|} \sum_{(x,y) \in S} xy, \\
\text{avg}(y) &= \frac{1}{|S|} \sum_{(x,y) \in S} y
\end{align*}
\]
Solution to normal equations:

\[ m = \frac{\text{avg}(xy) - \text{avg}(x) \cdot \text{avg}(y)}{\text{avg}(x^2) - \text{avg}(x)^2}, \]

\[ b = \text{avg}(y) - m \cdot \text{avg}(x) \]

What if \( \text{avg}(x^2) = \text{avg}(x)^2 \)?
For Dartmouth dataset:

\[ m = 0.751, \quad b = 0.067 \]

RMSE:

\[ \sqrt{\frac{1}{|S|} \text{sse}[m, b; S]} = 0.629 \]

(Recall standard deviation of college GPA is 0.75)
The diagram illustrates the relationship between high school GPA and college GPA. The data points represent individual student records, with each point indicating the high school GPA on the x-axis and the college GPA on the y-axis. The black line represents the Ordinary Least Squares (OLS) regression line, which predicts the expected college GPA based on the high school GPA.
Bivariate linear regression
Dartmouth dataset also includes SAT verbal percentiles.
Linear function of two variables $x_1$ and $x_2$:

$$f(x_1, x_2) = m_1 x_1 + m_2 x_2 + b$$

Problem: given a dataset $S$ from $\mathbb{R}^2 \times \mathbb{R}$, find (parameters of) a linear function $f(x_1, x_2) = m_1 x_1 + m_2 x_2 + b$ of minimal sum of squared errors

$$\text{sse}[m, b; S] = \sum_{(x_1, x_2, y) \in S} (m_1 x_1 + m_2 x_2 + b - y)^2$$
Normal equations: three linear equations in three unknowns \((m_1, m_2, b)\)

\[
\begin{bmatrix}
\text{avg}(x_1^2) & \text{avg}(x_1 x_2) & \text{avg}(x_1) \\
\text{avg}(x_2 x_1) & \text{avg}(x_2^2) & \text{avg}(x_2) \\
\text{avg}(x_1) & \text{avg}(x_2) & 1
\end{bmatrix}
\begin{bmatrix}
m_1 \\
m_2 \\
b
\end{bmatrix}
= 
\begin{bmatrix}
\text{avg}(x_1 y) \\
\text{avg}(x_2 y) \\
\text{avg}(y)
\end{bmatrix}
\]

Solve using elimination algorithm
Dartmouth dataset: $x_1 = \text{high school GPA}, \ x_2 = \text{SAT verbal percentile}$

$m_1 = 0.611, \ m_2 = 0.024, \ b = -0.639$

RMSE:

$$\sqrt{\frac{1}{|S|} \ \text{sse}[m_1, m_2, b; S]} = 0.603$$

(Recall standard deviation of college GPA is 0.75)
$m_1 \times (\text{high school GPA}) + m_2 \times (\text{SAT verbal percentile}) + b$
Linear algebra of ordinary least squares
(Homogeneous) linear function of $d$ variables $x = (x_1, \ldots, x_d)$ is parameterized by $d$-dimensional weight vector $w = (w_1, \ldots, w_d)$:

$$f_w(x) = w^T x$$

To handle inhomogeneous linear functions (i.e., affine functions), include an extra always-1 feature: $x_{d+1} = 1$

$$f_w(x) = w^T x$$

$$= (w_1 x_1 + \cdots + w_d x_d) + 1$$
Problem: given a dataset $S$ from $\mathbb{R}^d \times \mathbb{R}$, find $w \in \mathbb{R}^d$ of minimal sum of squared errors

$$\text{sse}[w; S] = \sum_{(x,y) \in S} (w^T x - y)^2$$

Method of solution: OLS
Matrix notation: let $S = ((x^{(i)}, y^{(i)}))_{i=1}^{n}$, and put

$$A = \begin{bmatrix}
\leftarrow \quad (x^{(1)})^T \quad \rightarrow \\
\vdots \\
\leftarrow \quad (x^{(n)})^T \quad \rightarrow 
\end{bmatrix}, \quad b = \begin{bmatrix}
y^{(1)} \\
\vdots \\
y^{(n)}
\end{bmatrix}$$

so

$$Aw = \begin{bmatrix}
w^T x^{(1)} \\
\vdots \\
w^T x^{(n)}
\end{bmatrix}, \quad Aw - b = \begin{bmatrix}
w^T x^{(1)} - y^{(1)} \\
\vdots \\
w^T x^{(n)} - y^{(n)}
\end{bmatrix}$$

Therefore

$$\|Aw - b\|^2 = \sum_{i=1}^{n}$$
$Aw \in CS(A) \text{ for every } w \in \mathbb{R}^d$
How many ways to write $\hat{b}$ as a linear combination of the columns of $A$?
Normal equations in matrix notation
Key fact: \( \text{CS}(A) \) and \( \text{NS}(A^T) \) are orthogonal complements
Summary:

- Normal equations: \((A^TA)w = A^Tb\)
- If \(\text{rank}(A) = d\), then solution is unique
- Else, infinitely-many solutions
- Common choice for tie-breaking: minimum norm solution

\[
\arg\min_{w \in \mathbb{R}^d} \|w\| \quad \text{s.t.} \quad (A^TA)w = A^Tb
\]
def learn(train_x, train_y):
    return np.linalg.pinv(train_x).dot(train_y)

def predict(params, test_x):
    return test_x.dot(params)
Statistical view of ordinary least squares
Normal linear regression model: Conditional distribution of $Y$ given $X = x$ is

$$N(w^T x, \sigma^2)$$

- $w$ and $\sigma^2$ are parameters of the model
- In this model, best possible MSE is $\sigma^2$
MLE in normal linear regression model

▶ Likelihood of $w$ and $\sigma^2$:

$$L(w, \sigma^2) = \prod_{(x,y) \in S} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - w^T x)^2}{2\sigma^2}\right)$$

▶ Log-likelihood:

$$\ln L(w, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{(x,y) \in S} (y - w^T x)^2 - \frac{|S|}{2} \ln(2\pi\sigma^2)$$

▶ In terms of $w$, maximizing log-likelihood is same as minimizing SSE!
Statistical inference (example)

Suppose you fit linear regression model to data, and find that $w \neq (0, \ldots, 0)$

How confident are you in this finding?
Generalization
- Suppose $S \overset{\text{i.i.d.}}{\sim} (X, Y)$
- OLS gives minimizer of empirical risk (for square loss, among linear functions)

$$\hat{\text{Risk}}[w] = \frac{1}{n} \sum_{(x,y) \in S} \text{loss}_{\text{sq}}(w^T x, y)$$

But we actually care about the (true) risk

$$\text{Risk}[w] = \mathbb{E}[\text{loss}_{\text{sq}}(w^T X, Y)]$$

- Is empirical risk a good estimate of (true) risk?
  - Usually only if $|S|$ is sufficiently large
Extreme example: $d = 1, |S| = 2, \widehat{\text{Risk}}[w] = 0$