Data Mining

CS57300
Purdue University

Jan 11, 2018

Bruno Ribeiro
• Regression
• Posteriors
• Working with Data
Linear Regression: Review
Linear Regression (use A)

- Interpolation
  (something is missing)
- \((x_1, \ldots, x_t)\)
- \((y_1, \ldots, y_t)\)

\[ y = \alpha x \]
\[ x_k \approx \frac{y}{\alpha} \]
Auto-regression: Predicting Next Value After $t$ Steps
Linear Regression (use B)

Similar problem to linear regression:
express unknowns as a linear function of knowns
Predictions from High-Dimensional Historical Data

\[ y[t \times 1] = X[t \times w] a[w \times 1] \]

- Over-constrained problem
  - \( a \) is the vector of the regression coefficients
  - \( X \) has the \( t \) values of the \( w \) independent variables. These independent variables can mix user characteristics with a window of past observations
  - \( y \) has the \( t \) values of the dependent variable
Looking Into Multiplication

Predicting corn prices over time…

\[ y_{[t \times 1]} = X_{[t \times w]} a_{[w \times 1]} \]

\[
\begin{bmatrix}
X_{11}, X_{12}, \ldots, X_{1w} \\
X_{21}, X_{22}, \ldots, X_{2w} \\
\vdots \\
X_{t1}, X_{t2}, \ldots, X_{tw}
\end{bmatrix}
\times
\begin{bmatrix}
a_1 \\
a_2 \\
\vdots \\
a_w
\end{bmatrix}
= 
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_t
\end{bmatrix}
\]
How to Estimate $a$?

- $a = (X^T \cdot X)^{-1} \cdot (X^T \cdot y)$

$X^+ = (X^T \cdot X)^{-1} \cdot X^T$ is the Moore–Penrose pseudoinverse

Or: $a = X^+ y$

$a$ is the vector that minimizes the Root Mean Squared Error (RMSE) of $(y - X \cdot a^T)$
Details: Least Squares Optimization

• Least squares cost function:

\[ C = \frac{1}{2} \sum_{i=1}^{t} (y_i - x_i^T a)^2 = \frac{1}{2} (y - Xa)^T (y - Xa) \]

• Find \( a \) that minimizes cost \( C \)

\[
\frac{\partial C}{\partial a} = \frac{1}{2} \frac{\partial}{\partial a} (y - Xa)^T (y - Xa) = -(y - Xa)^T X
\]

• Optimal value at:

\[
\frac{\partial C}{\partial a} = 0 \implies X^T y = X^T X a \implies a = (X^T X)^{-1} X^T y
\]
How to Estimate $a$?

• $a = (X^T \cdot X)^{-1} \cdot (X^T \cdot y)$

$X^+ = (X^T \cdot X)^{-1} \cdot X^T$ is the Moore–Penrose pseudoinverse

Or: $a = X^+ y$

$a$ is the vector that minimizes the Root Mean Squared Error (RMSE) of $(y - X \cdot a^T)$

Problems:

Matrix $X$ grows over time & needs matrix inversion

• $O(t \cdot w^2)$ computation

• $O(t \cdot w)$ storage
Recursive Least Squares

At time $t$ we know $X_t = (x_1, \ldots, x_t)$, $y_t = (y_1, \ldots, y_t)$

Least squares is solving

$$\argmax_{\mathbf{a}} \| \mathbf{a}^T X_t - y_t \|^2$$

which gives

$$\mathbf{a}^* = X^+ y$$

where $X^+ = (X^T \cdot X)^{-1} \cdot X^T$

Let

$$\Phi_t = X_t^T X_t \quad \theta_t = X_t^T y_t$$

Then $\Phi_{t+1}^{-1}$ is

$$\Phi_{t+1}^{-1} = (\Phi_t + x_{t+1} x_{t+1}^T)^{-1} = \Phi_t^{-1} - \frac{\Phi_t^{-1} x_{t+1} x_{t+1}^T \Phi_t^{-1}}{1 + x_{t+1}^T \Phi_t^{-1} x_{t+1}}$$
Recursive Least Squares Algorithm

\[
\Phi_{t+1}^{-1} = \Phi_t^{-1} - \frac{\Phi_t^{-1} x_{t+1}^T x_{t+1} \Phi_t^{-1}}{1 + x_{t+1}^T \Phi_t^{-1} x(t + 1)}
\]

\[
\theta_{t+1} = \theta_t + x_{t+1}^T y_{t+1}
\]

\[
a_{t+1} = \Phi_{t+1}^{-1} \theta_{t+1}
\]
Exponentially Weighted Recursive Least Squares Algorithm

for $\lambda > 1$

$$
\Phi_{t+1}^{-1} = \frac{1}{\lambda} \Phi_t^{-1} - \frac{1}{\lambda^2} \frac{\Phi_t^{-1} x_{t+1}^T x_{t+1} \Phi_t^{-1}}{1 + x_{t+1}^T \Phi_t^{-1} x(t + 1)}
$$

$$
\theta_{t+1} = \lambda \theta_t + x_{t+1}^T y_{t+1}
$$

$$
a_{t+1} = \Phi_{t+1}^{-1} \theta_{t+1}
$$
Comparison

**Original Least Squares**
- Needs large matrix (growing in size) $O(t \times w)$
- Costly matrix operation $O(t \times w^2)$

**Recursive LS**
- Need much smaller, fixed size matrix $O(w \times w)$
- Fast, incremental computation $O(1 \times w^2)$
- no matrix inversion
Posteriors
Finding (Real) Patterns in Data

- **Data shows**: rural counties have the highest average mortality rates. But rural counties also have small populations.

- Why rural counties have the highest rates of cancer?
  - **A**: Small sample variance
  - **Solution?**
  - $P[\text{cancer rate} | \text{data}]$
Posteriors (e.g. Representing Fractions)

• Consider creating a model that predicts if a soccer striker will score a goal in a game
• Data includes \textit{no. shots on goals} and \textit{no. goals} during the player’s career

• Problems with absolute values:
  • number of goals
    (older players have larger values than young players)
  • no. shots on goals
    (does not reflect rate of shot $\rightarrow$ goal conversion)

• Feature: \% shots on goal resulting in goal
  • Alice (Novice): 1 out of 1 $\rightarrow$ 100\%
  • Bob (Senior): 300 out of 1000 $\rightarrow$ 30\%

• Solution? (same problem as in the previous slide (cancer rate). The solution is also Bayesian.)
Reaching (Real) Conclusions with Data

• Whatever conclusion using ”cold data”, there are always assumptions

• Example: Simpson’s Paradox (a.k.a. Yule–Simpson effect)

  • At Berkeley, women applicants overall have lower acceptance rate than men

  • But if you look at every department women are accepted at the same rate as men

• How? What was our wrong assumption?

  • $O_i$ is a random variable that denotes candidate $i$ gets an offer from Berkeley.

  • $O_{i,j}$ is a random variable that denotes candidate $i$ gets an offer from department $j$ at Berkeley.

  • $A_{i,j}$ is a random variable that denotes candidate $i$ has applied to department $j$.

\[
P[O_i] = \sum_j P[O_{i,j}] = \sum_j P[O_{i,j} | A_{i,j}] P[A_{i,j}]
\]

Incorrect assumption: $P[A_{i,j}]$ is the same for men and women
Working with Data
Data Issues

• How to represent the data plays a huge role in the question we can ask and the answers we get
  • It influences similarity metrics
  • It influences the data models we can use

• Data biases (last class)

• Missing data also plays a huge role

• And the problem of outliers in the data
  • Outlier chicken-and-egg problem:
    • Outliers may skew decisions
    • But defining what constitutes an outlier requires deciding a model that describes the “normal” data
    • Deciding such a model requires fitting the data, which may fit the outliers
Missing values

• Reasons for missing values
  • Information is not collected (e.g., people decline to give their age)
  • Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)

• Ways to handle missing values
  • Eliminate entities with missing values
  • Estimate attributes with missing values
  • Ignore the missing values during analysis
  • Replace with all possible values (weighted by their probabilities)
  • Impute missing values
Duplicate Data

• Data set may include data entities that are duplicates, or almost duplicates of one another
  • Major issue when merging data from heterogeneous sources
  • Example: same person with multiple email addresses
• Sometimes “duplication” happens (different users, same features).
  • Issues with some models.
• Data cleaning
  • Finding and dealing with duplicate entities
  • Finding and correcting measurement error
  • Dealing with missing values