Supervised Learning in Data Science

Data Analysis and Policy

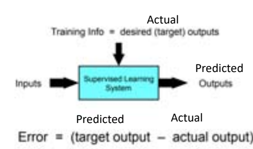
- Google wants to hire new software engineers that increase productivity as much as possible. It collects the following data on 2017 hires:
  - Productivity increase contribution
  - Cumulative college GPA
  - Number of undergraduate internships held
  - Extracurricular activity participation
  - Number of CS courses taken
- What can Google do with this data to inform this year's hiring decisions?

Data Analysis and Policy

- Causal Inference – What effect does a specific input variable have on the output?
- Relevance – Which input variables actually matter in determining the output?
- Prediction – Knowing all the input values, what can we expect the output to be?
  - Primary goal of supervised learning

Supervised Learning

- Have a training data set, know what output given inputs should look like
- Want to teach computer to give correct output given inputs



The Regression Problem

- Theory posits that some quantitative output is a function of input and random error:

$$Y = f(X) + \epsilon$$

- Given a set of observations $(X^{(1)}, Y^{(1)}), (X^{(2)}, Y^{(2)}), \ldots, (X^{(n)}, Y^{(n)})$ try to create predictor function $\hat{f}(X)$ which produces an output as close to Y as possible on a new input value of X

Linear Predictor Function

- With inputs $X_1 \ldots X_k$, actual output Y, linear predictor function takes the form

$$\hat{f}(X) = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \cdots + \widehat{\beta_k}X_k$$

where $\widehat{\beta_0} \ldots \widehat{\beta_k}$ are user determined parameters

- Single variable case:

$$\hat{f}(X) = \widehat{\beta_0} + \widehat{\beta_1}X$$

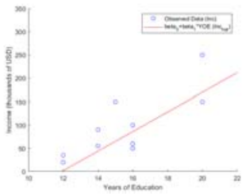
- Note: predicted output $\hat{Y} := \hat{f}(X)$

Linear Predictor Function

- E.g. income based on years of education

$Inc = f(YOE) + \epsilon$

$\widehat{Inc} = \widehat{\beta_0} + \widehat{\beta_1} YOE$



Cost Function

- How to evaluate performance of $\hat{f}$ i.e. measure goodness-of-fit?

  Mean squared error

- Cost function: $C(\widehat{\beta_0}, \widehat{\beta_1}) = \frac{1}{2n} \sum_{i=1}^{n} \left(\hat{f}(X^{(i)}) - Y^{(i)}\right)^2$
- Goal: choose $\widehat{\beta_0}, \widehat{\beta_1}$ to minimize $C(\widehat{\beta_0}, \widehat{\beta_1})$, maximize goodness of fit

Gradient Descent

- How to "learn" from cost function and reduce it? Correction factors
  - $\delta_0 = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}(X^{(i)}) - Y^{(i)}\right)$
  - $\delta_1 = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}(X^{(i)}) - Y^{(i)}\right) X^{(i)}$
- Reassign $\widehat{\beta_0}, \widehat{\beta_1}$ such that:

$$\widehat{\beta_0} \coloneqq \widehat{\beta_0} - \alpha\, \delta_0$$
$$\widehat{\beta_1} \coloneqq \widehat{\beta_1} - \alpha\, \delta_1$$

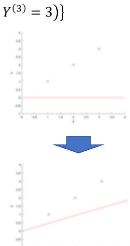
- α is the learning rate of the gradient descent algorithm, user determined
  - Too slow if too small; could overshoot if too large

Gradient Descent

- Algorithm:
  1. Initialize $\widehat{\beta_0}, \widehat{\beta_1}$
  2. Calculate δ_0, δ_1 and for $j = 0,1$ reassign $\widehat{\beta_j} \coloneqq \widehat{\beta_j} - \alpha \delta_j$
  3. Repeat step 2 until stopping condition reached
- Result: $\hat{f}(X)$ which predicts Y with minimal cost!

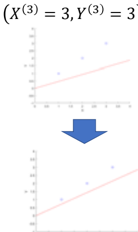
Gradient Descent Example

- $TS \coloneqq \{(X^{(1)} = 1, Y^{(1)} = 1), (X^{(2)} = 2, Y^{(2)} = 2), (X^{(3)} = 3, Y^{(3)} = 3)\}$
- Initialize $\widehat{\beta_0} = 0, \widehat{\beta_1} = 0$, such that $\hat{f}(X) = 0 + 0X$
  - $\widehat{\beta_0}$ already optimal, just optimize by $\widehat{\beta_1}$
- $C(\widehat{\beta_0}, \widehat{\beta_1}) = 2.3333$
- Let $\alpha = 0.1$
- $\delta_1 = -4.6667$
- $\widehat{\beta_1} \coloneqq 0 - 0.1 * (-4.6667) = 0.4667$



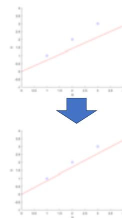
Gradient Descent Example

- $TS \coloneqq \{(X^{(1)} = 1, Y^{(1)} = 1), (X^{(2)} = 2, Y^{(2)} = 2), (X^{(3)} = 3, Y^{(3)} = 3)\}$
- $C(\widehat{\beta_0}, \widehat{\beta_1}) = 0.6636$
- Let $\alpha = 0.1$
- $\delta_1 = -2.4887$
- $\widehat{\beta_1} \coloneqq 0.4667 - 0.1 * (-2.4887) = 0.7156$



Gradient Descent Example

- $TS \coloneqq \{(X^{(1)} = 1, Y^{(1)} = 1), (X^{(2)} = 2, Y^{(2)} = 2), (X^{(3)} = 3, Y^{(3)} = 3)\}$
- $C(\widehat{\beta_0}, \widehat{\beta_1}) = 0.1887$
- Let $\alpha = 0.1$
- $\delta_1 = -1.3272$
- $\widehat{\beta_1} \coloneqq .7156 - 0.1 * (-1.3272) = 0.8483$



Solving for a Multivariate Predictor

- Linear predictor function: $\hat{f}(X_1, X_2 \dots, X_k) = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \widehat{\beta_2} X_2 + \cdots + \widehat{\beta_k} X_k$
- Goal: minimize $C(\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2} \dots, \widehat{\beta_k}) = \frac{1}{2n} \sum_{i=1}^{n} (\hat{Y}^{(i)} - Y^{(i)})^2$
- δ_0 unchanged, $= \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}^{(i)} - Y^{(i)})$
- For $j = 1, 2 \dots, k$, $\delta_j = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}^{(i)} - Y^{(i)}) X_j^{(i)}$

Solving for a Multivariate Predictor

- Algorithm:
  1. Initialize $\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2} \dots, \widehat{\beta_k}$
  2. Calculate $\delta_0, \delta_1, \dots, \delta_k$ and for $j = 0, 1, \dots, k$ reassign $\widehat{\beta_j} \coloneqq \widehat{\beta_j} - \alpha \delta_j$
  3. Repeat step 2 until stopping condition reached

Overfitting

- Occurs when estimation method adheres too closely to training data, picks up random noise and poorly expresses actual structure of data
- Can result from too many variables being used in linear regression model
  - E.g. If one is trying to predict an individuals income, birthday probably isn't relevant, but doesn't hurt to consider it

Omitted Variable Bias

- Occurs when important factors aren't considered in estimation method
  - i.e. when relevant variables aren't included in linear regression model
- Results in certain variables being considered more or less important than they really are
  - E.g. if innate ability explains both years of education and income, education can act as a proxy for ability in absence of data on the latter, but will be weighted too highly

Bias in Data

- Occurs when training data set has structural differences from overall population
- E.g. if predicting height while only looking at schoolchildren, age is a highly descriptive variable; however, less relevant for overall population