Day 3, 9/14/18

Problem: adaptively select learning rate α

Intuition: large α is good, especially in early stages of optimization, as we advance faster towards the optimum.
But we can overshoot. If we overshoot, we backtrack (i.e., start over, with a reduced α)

Algo: line search (replaces the highlighted part of batch/stochastic gradient descent

- Input: $\theta^{(t)}$, $J(\theta)$, $\nabla J(\theta)$
- Parameters: α_{max}, $\tau \in [0.5, 0.9]$, tolerance, maxBacktrack
- $\alpha \leftarrow \alpha_{max}$
- $\theta \leftarrow \theta^{(t)}$
- $obj \leftarrow J(\theta)$
- Repeat until maxBacktrack {
  - For $j = 1, \dots, p$ {
    $\theta_j \leftarrow \theta_j + \alpha \frac{d}{d\theta_j} \sum_{i=1}^{N} (y_i - x_i^T \theta^{(t)})^2$ [batch version]
  - } if $J(\theta_j) < obj - tolerance$, then break [improved $J(\theta)$]
    else $\alpha \leftarrow \tau\alpha$ [backtrack]
- }
- If maxBacktrack reached, return $\{\theta^{(t)}, \alpha = 0\}$
  else return $\{\theta, \alpha\}$

Practical considerations in Linear regression

Empirical measures of association when $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & Cov(Y,X) \\ Cov(Y,X) & \sigma_Y^2 \end{pmatrix}\right)$

Pearson correlation

Population: $Y = \beta_0 + \beta_1 X + \varepsilon$

data collection

Data: $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$

conclusions

$$\rho = \frac{Cov(Y,X)}{\sigma_Y \sigma_X} = \frac{E\{(Y-\mu_Y)(X-\mu_X)\}}{\sigma_Y \sigma_X}$$

$$E\{Y|X\} = \underbrace{(\mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X})}_{\beta_0} + \underbrace{\rho \frac{\sigma_Y}{\sigma_X}}_{\beta_1} X$$

$$r = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}}$$

$$r = \hat{\beta}_1 \frac{\sqrt{\sum (x_i - \bar{x})^2}}{\sqrt{\sum (y_i - \bar{y})^2}}$$

r: Pearson correlation coef

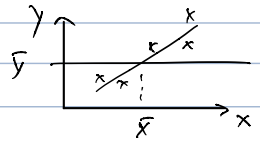
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

least squares estimate of slope

Conclusion: empirical summaries mirror population summaries. However,

correlation does not extend to multiple predictors

Coefficient of multiple determination R^2



$$\sum_{i=1}^{N}(y_i-\bar{y})^2 = \sum_{i=1}^{N}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2$$

$$\sum_{i=1}^{N}(y_i-\bar{y})^2 = \sum_{i=1}^{N}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2 \qquad [\text{the cross-product} = 0]$$

$SSTO$ — sum of squares total; SSE — some of squares of the error; SSR — some of squares of regression

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO},$$ i.e. the proportion of total variation in y that is explained by its linear association with X

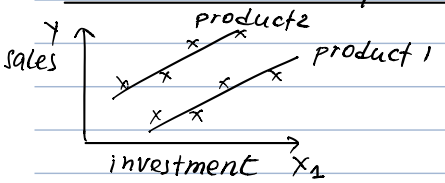
Plugging in the expression for $\hat{\delta}_1$:

$$SSR = \sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2 = \sum_{i=1}^{N}\left(\hat{\delta}_0+\hat{\delta}_1 x_i - \frac{1}{N}\sum_{j=1}^{N}\hat{\delta}_0 - \frac{1}{N}\sum_{j=1}^{N}\hat{\delta}_1 x_i\right)^2 = \hat{\delta}_1^2\sum_{i=1}^{N}(x_i-\bar{x})^2$$

$$R^2 = \frac{\hat{\delta}_1^2\sum_{i=1}^{N}(x_i-\bar{x})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} = r^2,$$ i.e. in linear regression with one predictor, the coefficient of multiple determination is the square of Pearson correlation coefficient

R^2 extends to multiple predictors; r does not. Neither quantify the linearity of the association

Qualitative predictors



x_1 : investment

$$x_2 = \begin{cases} 0, & \text{product 1} \\ 1, & \text{product 2} \end{cases}$$

$$y_i = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \varepsilon_i$$

Product 1: $y_i = \delta_0 + \delta_1 x_1 + \varepsilon_i$

Product 2: $y_i = (\delta_0+\delta_2) + \delta_1 x_1 + \varepsilon_i$

} Different intercepts, same slope, same variance

$$y_i = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_1 x_2 + \varepsilon_i$$

Product 1: $\delta_0 + \delta_1 x_1 + \varepsilon_i$

Product 2: $(\delta_0+\delta_2) + (\delta_1+\delta_3)x_1 + \varepsilon_i$

} Different intercepts, Different slopes, same variance

3 products:

$$x_2 = \begin{cases} 0, & \text{product 1} \\ 1, & \text{product 2} \\ 0, & \text{product 3} \end{cases} \qquad x_3 = \begin{cases} 0, & \text{product 1} \\ 0, & \text{product 2} \\ 1, & \text{product 3} \end{cases}$$

Note: coding $x_2 = \begin{cases} 0 & \text{product 1} \\ 1 & \text{product 2} \\ 2 & \text{product 3} \end{cases}$ is inappropriate, because it assumes that the difference in intercepts between products 3 and 1 is twice as big as between products 2 and 1

Simpson's paradox: change of association between two variables when condition on another variable