Review

Tianhang Zheng

https://tianzheng4.github.io

Supervised Learning

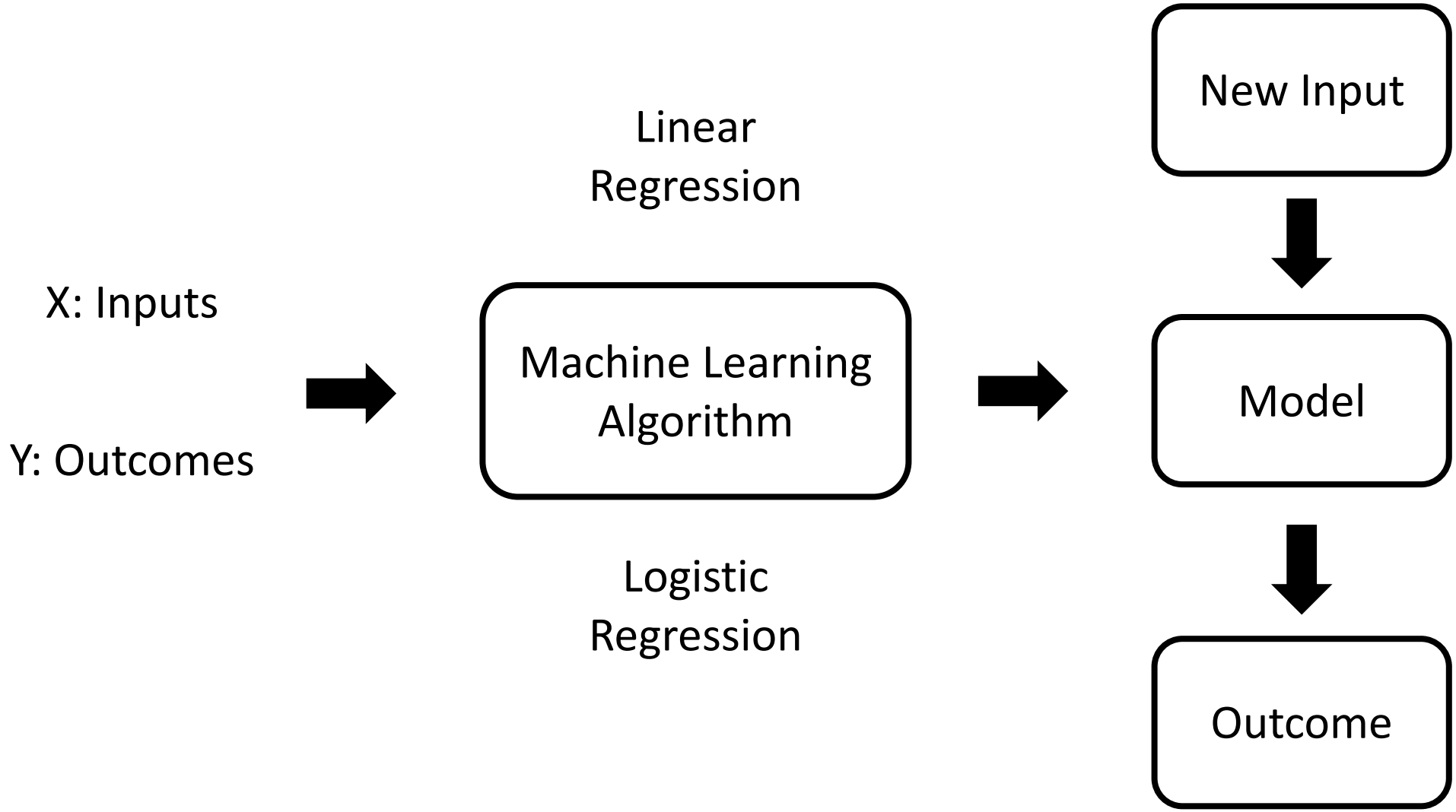
Vector/Matrix/Tensor predictor X (also called inputs, regressors, covariates, features, independent variables)

Outcome Y (also called dependent variable, response, target)

Objectives:

1. Accurately predict the outcomes of unseen test cases
2. Understand which inputs affect the outcome, and how
3. Assess the quality of our predictions and inferences

Supervised Learning

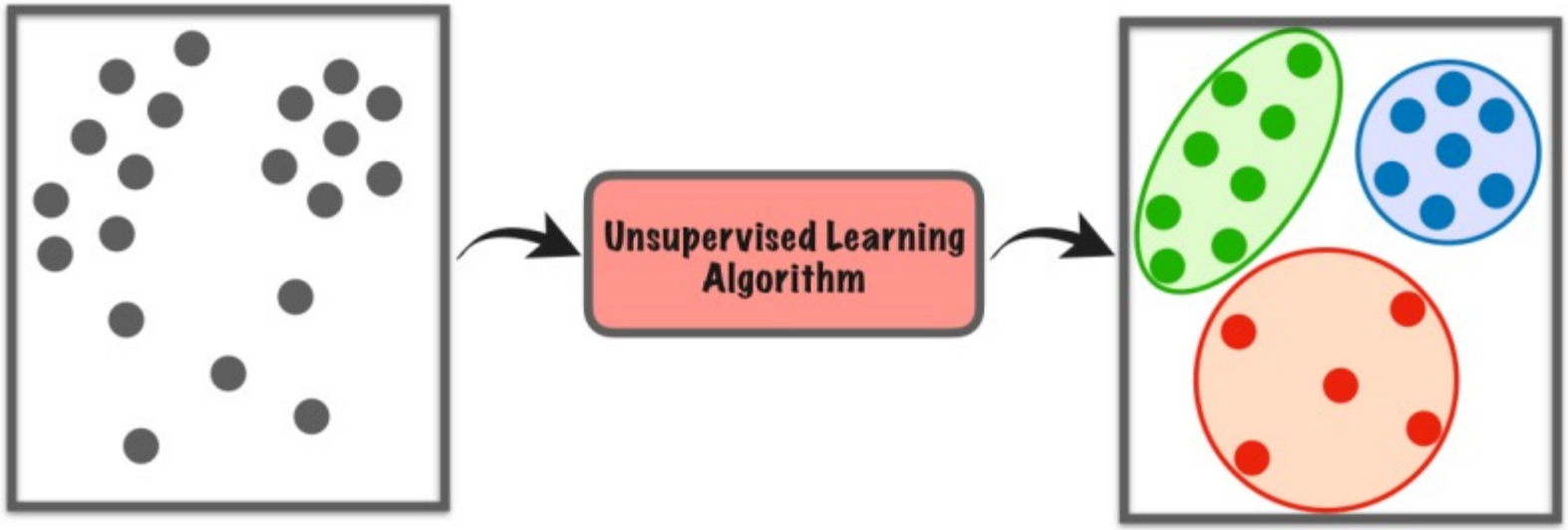


Unsupervised Learning

No outcome variable, just a set of predictors (features) measured on a set of samples.

Objectives:

1. Find groups of samples that behave similarly
2. Find the most important sets of features



How to assess a model

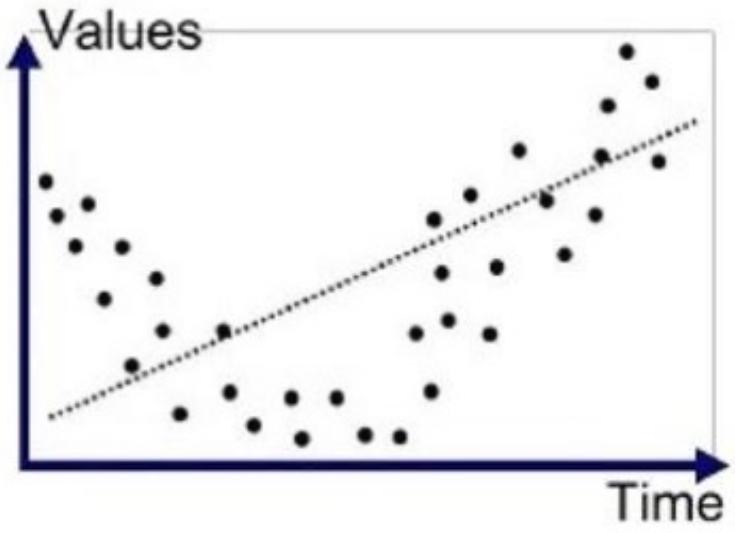
Prediction Error (regression problems)

Prediction Accuracy (classification problems)

Model Variance

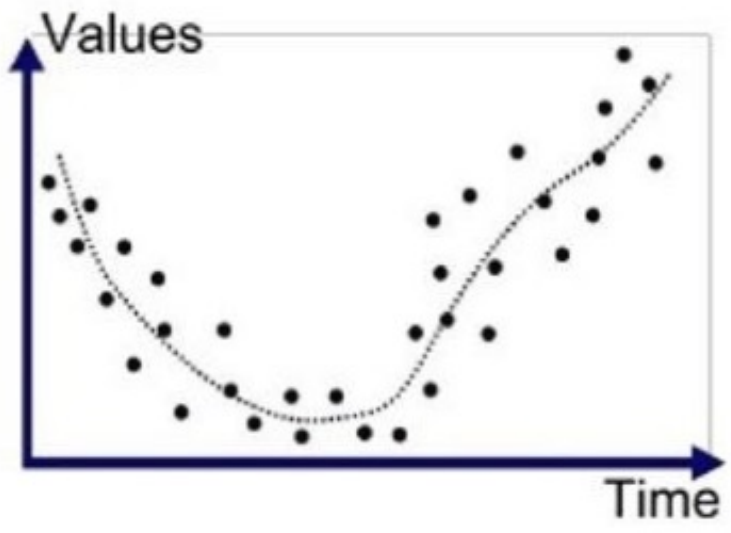
Interpretability

Goodfit, Underfit and Overfit

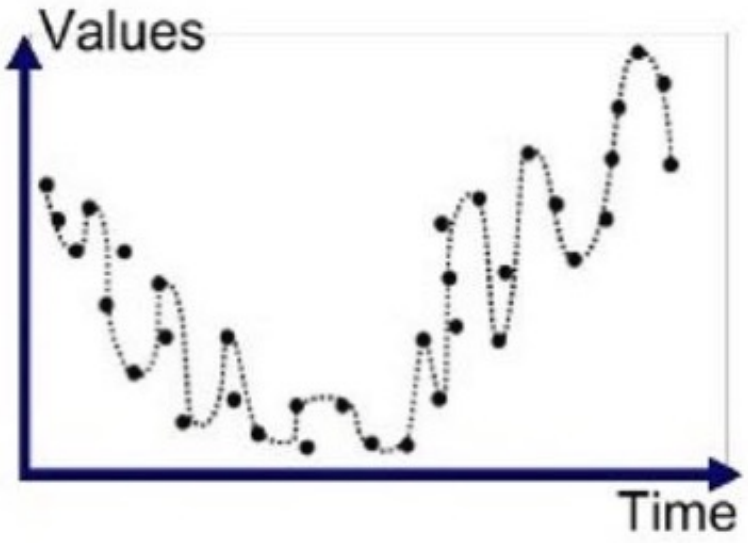


Underfitted

High bias



Good Fit/Robust



Overfitted

High variance

Bias and Variance (Corrections)

$$E_D\left[\left(Y - g(X)\right)^2\right] = bias^2 + variance + \sigma^2$$

Bias: $E_D[g(X;D)] - f(X)$

Variance: $\mathrm{E_D}[(\mathrm{E}_D[g(X;D)] - g(X;D))^2]$

} Depends on model complexity

Irreducible error: σ

Simple Linear Regression

Linear Regression with a single predictor (Assume the ideal model is a linear function)

$$Y = \beta_0 + \beta_1 X + \epsilon$$

β_0 is called intercept and β_1 is called slope, which are two parameters.

ϵ is the error term: $\epsilon \sim N(0, \sigma^2)$

Least Squares Method

$$\min_{\hat{\beta}_0,\hat{\beta}_1} RSS = \sum_i e_i^2 = \sum_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Take the derivative and set the derivative as 0

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Confidence Level

$\hat{\beta} \sim N(\beta, SE^2(\hat{\beta}))$ means that $\beta \sim N(\hat{\beta}, SE^2(\hat{\beta}))$

A 95% confidence interval is defined as a range of values with 95% probability, and the interval for the least square method is

$$[\hat{\beta} - 2SE(\hat{\beta}), \hat{\beta} + 2SE(\hat{\beta})]$$

There is 95% probability that this interval contains the true β

Prediction Error

The residual sum of squares (RSS)

$$RSS = \sum_i e_i^2 = \sum (y_i - \hat{y}_i)^2$$

The residual standard error (RSE)

$$RSE = \sqrt{\frac{1}{n-2} RSS}$$

R-Squared

The proportion of the variance that can be explained by a model

$$R^2 = 1 - \frac{RSS}{TSS}$$

TSS is the total sum of squares (total variance of y)

$$TSS = \sum_i (y_i - \bar{y})^2$$

Multiple Linear Regression

Linear Regression with multiple predictors (Assume the ideal model is a linear function)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

β_0 is interpreted as the average effect of one unit increase in X_i on Y

Least Square Method (Solved by Matrix)

$$y = \begin{bmatrix} 6 \\ 11 \\ 4 \\ 3 \\ 5 \\ 9 \\ 10 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 4 & 5 & 4 \\ 1 & 7 & 2 & 3 \\ 1 & 2 & 6 & 4 \\ 1 & 1 & 9 & 6 \\ 1 & 3 & 4 & 5 \\ 1 & 7 & 3 & 4 \\ 1 & 8 & 2 & 5 \end{bmatrix}$$

Estimation:

$$b = (X'X)^{-1}X'y$$

$$b = [\hat{\beta}_0, \hat{\beta}_1 \ldots, \hat{\beta}_3]$$

$$y = Xb + e$$

Hypothesis Testing

$$H_0: \beta_j = 0$$

We can use t-statistics $$t = \frac{\hat{\beta}_j - 0}{\widehat{SE}(\hat{\beta}_j)}$$

$$\widehat{\boldsymbol{\beta}} = \left[\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right] \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\mathbf{t}\mathbf{X})^{-1})$$

P value

A p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis (H_0) is correct.

$$\text{P-value} = P[T > |t|]$$

If p-value is large, we tend to accept H_0. Otherwise, we tend to reject it.

Logistic Regression

$$p(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

$p(Y = 1|X)$ always has values between 0 and 1

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \mathrm{X}$$

Maximum Likelihood Method

Commonly used for parameter estimation of logistic regression

Assume that the predictors are independent

$$p(y|x) = \prod_{i:y_i=1} p_i \prod_{i:y_i=0} (1 - p_i) \qquad p_i = p(y_i = 1|x_i)$$

This likelihood characterizes the conditional probability of the observed data

Multi-Class Logistic Regression

A linear function for each class

$$p(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X}}{\sum_{l=1}^{K} e^{\beta_{0l}+\beta_{1l}X}}$$

Multi-class logistic regression is also called multinomial regression

K Nearest Neighbors

A non-parametric supervised learning algorithm for classification

Assign the label of x based on a majority vote mechanism

Select the k training points that are nearest to the target point x

Assign the majority label for the k training points as the label of x

K-Fold Cross-validation

Randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other K − 1 parts (combined), and then obtain predictions for the left-out kth part.

This is done in turn for each part k = 1,2,...K, and then the results are combined.

K-Fold Cross-validation

1	2	3	4	5
Validation	Train	Train	Train	Train

Denote K parts by $C_1, C_2, C_3 \ldots, C_K$

$$\mathrm{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{MSE}_k \qquad \mathrm{MSE}_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$$

Leave-One Out Cross-Validation (LOOCV)

K=n K-fold cross validation

LOOCV is sometimes useful, but the estimates are highly correlated, so the average has high variance

$$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y)$$

A better choice is K=5 or 10

BootStrap

We obtain distinct datasets by repeatedly sampling observations from the original dataset with replacement.

Each “bootstrap dataset” is created by sampling with replacement and is the same size as our original dataset.

Use the bootstrap to get an estimate of a parameter

BootStrap

Primarily used to obtain standard errors of an estimate.

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \bar{\hat{\alpha}}^{*}\right)^2}.$$

Provide approximate confidence intervals for a population parameter.

BootStrap to Estimate Prediction Error?

To estimate prediction error using the bootstrap, we could think about using each bootstrap dataset as our training sample, and the original sample as our validation sample.

But there is large overlap between training and validation set, which will underestimate the error

Final Exam

- Three concept questions
- Two calculation questions