TIM 245 Lecture 7 (4/24/17)

Agenda

1) General process for supervised learning problems

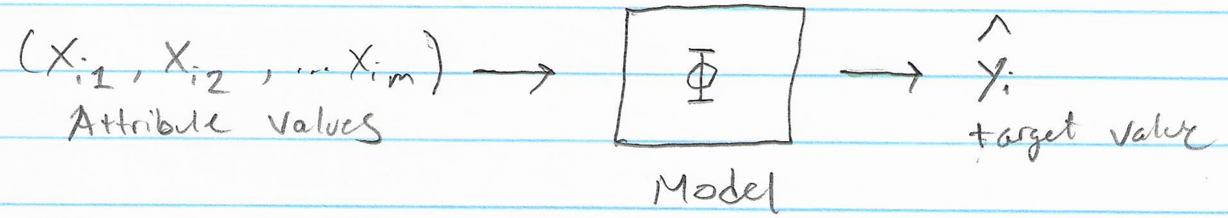
2) Evaluation of classification and prediction models

3) Linear regression and introduction to prediction models

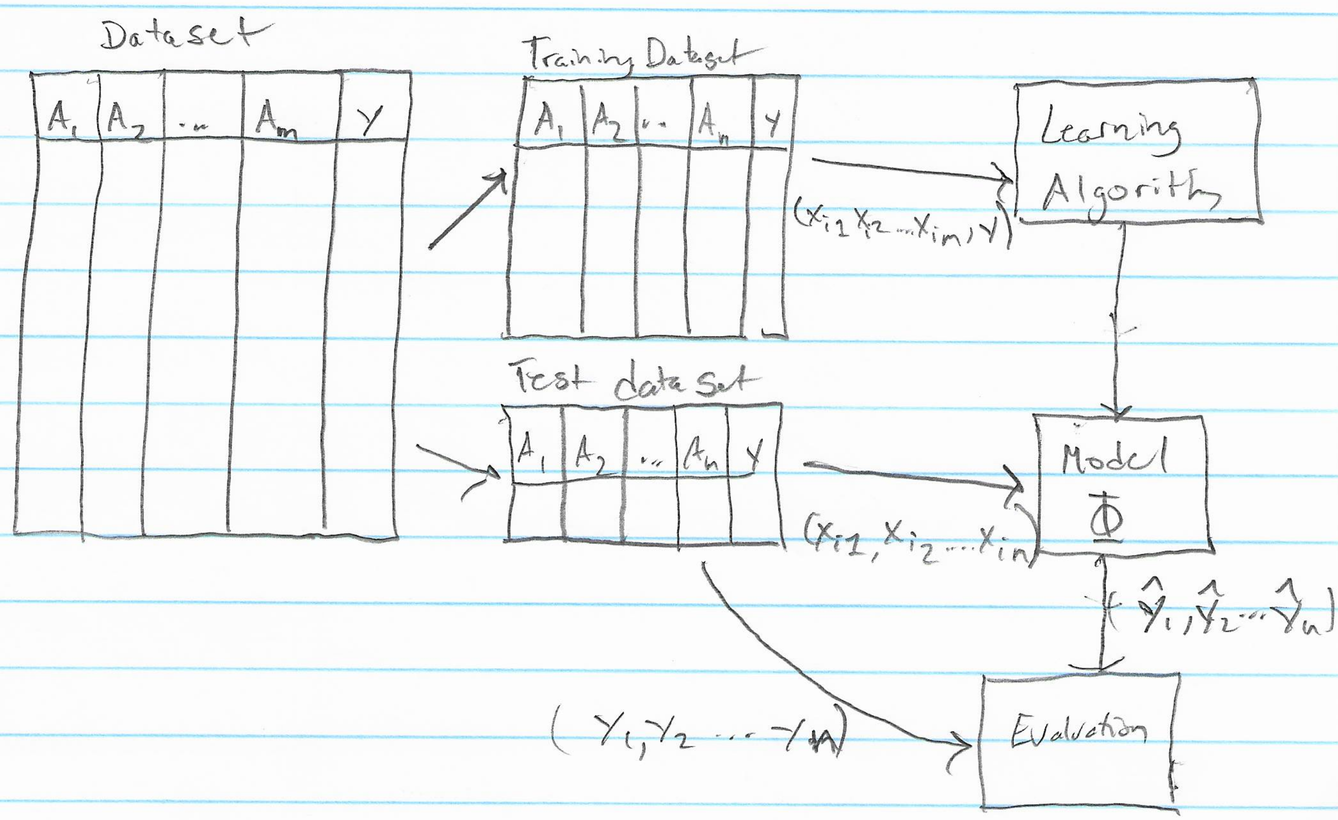
4) Work on project (Time permitting)

(1) General Process for Supervised learning problems

Objective: Create a function or model that can map an input set of attributes to a target output value



The model, Φ, is created through a supervised learning process:



Two general applications of supervised learning

Descriptive Modeling

use Φ to understand relationships in the historical data

(Focus is interpretability)

Predictive Modeling

use Φ to predict y_i for a new instance $(x_{i1}, x_{i2}, \ldots, x_{ik})$

(Focus is accuracy)

Key issue in Descriptive Modeling: Correlation vs Causation

Four possible cases for $\Phi: X \to Y$

1) Causal Link

$$\left.\begin{array}{l} X \to Y \\ Y \to X \end{array}\right\} \text{No sure which one}$$

2) Hidden Cause

$$\left.\begin{tikzcd} & Z \arrow[dl] \arrow[dr] & \\ X & & Y \end{tikzcd}\right\} \text{Z is not in the data set}$$

3) Confounding Factor

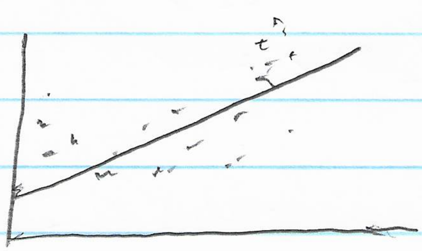
$$\begin{tikzcd} Z \arrow[dr] & \\ & Y \\ X \arrow[ur] & \end{tikzcd}$$

4) Coincidence (Noise)

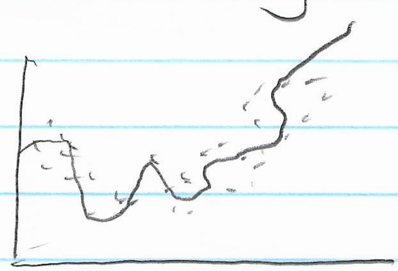
X Y

Typically addressed through:
- Careful with attribute selection
- Consulting with subject matter experts
- Randomized experiments (A/B testing)

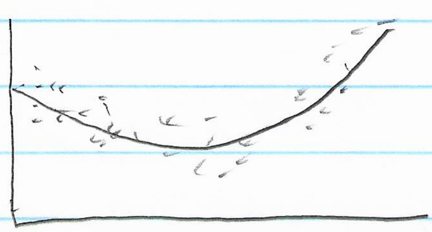
Key issue in Predictive Modeling: overfitting vs underfitting



underfitting



overfitting



Just right

Typically addressed through:

- Model regularization (Lasso, Ridge)
- Careful evaluation of the accuracy
- Cross validation
- Ensemble methods (Bagging, Boosting)

② Evalvation of Classification and prediction models

Classification model are evalvated using a confusion matrix

		Predicted: Award	Predicted: No Award
Actual	Award	True Positives	False Negative
	No Award	False Positive	True Negatives

Measures: Accuracy, Precision, Recall, f-measure, g-mean

Example: $\text{Accuracy} = \dfrac{TP + TN}{TP + TN + FP + FN}$

Prediction Models are evalvated based on the error

$$e_i = y_i - \hat{y}_i \quad \text{(residual)}$$

Measures: Sum of Squared Error (SSE),
Mean Absolute Error (MAE),
Mean Average Percent Error (MAPE)

Example: $SSE = \sum_{i=n}^{n} (y_i - \hat{y}_i)^2$

③ Linear Regression

three general types of relationships between X and Y

<u>Positive Correlation</u>: values move in same direction

↑X ↑Y or ↓X ↓Y

<u>Negative Correlation</u>: Values move in opposite directions

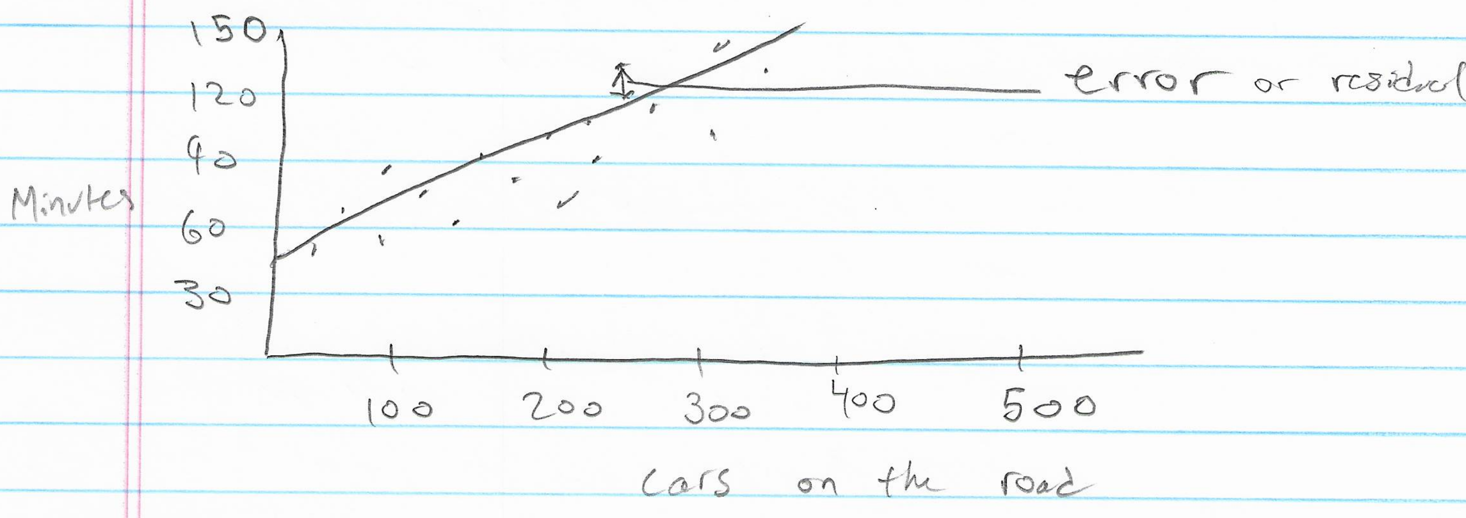
↑X ↓Y or ↓X ↑Y

<u>No Correlation</u>

↑X Y or ↓X Y or X↑Y or X↓Y

Linear regression estimates the type and strength of linear relationships in the training data set, i.e. best straightline fit

Example: HWY 17 Commute times vs number of cars on the road



General form of a linear regression model

$$\hat{y} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im}$$

dependent variable   intercept   Coefficients   independent Variables

Coefficients indicate the strength and type of the relationship:

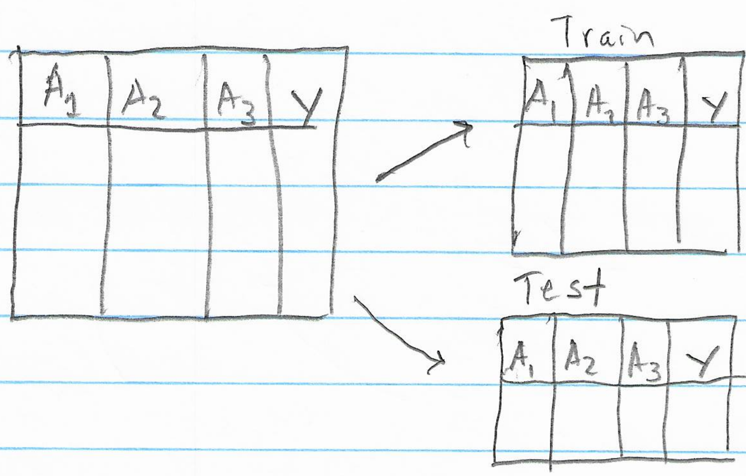
Magnitude : Strength of the relationship
≈ 0 : no correlation
Sign : type of relationship
$+$: positively correlated
$-$: negatively correlated

<u>Simple Linear Regression</u>: linear regression with only a single independent variable or attribute

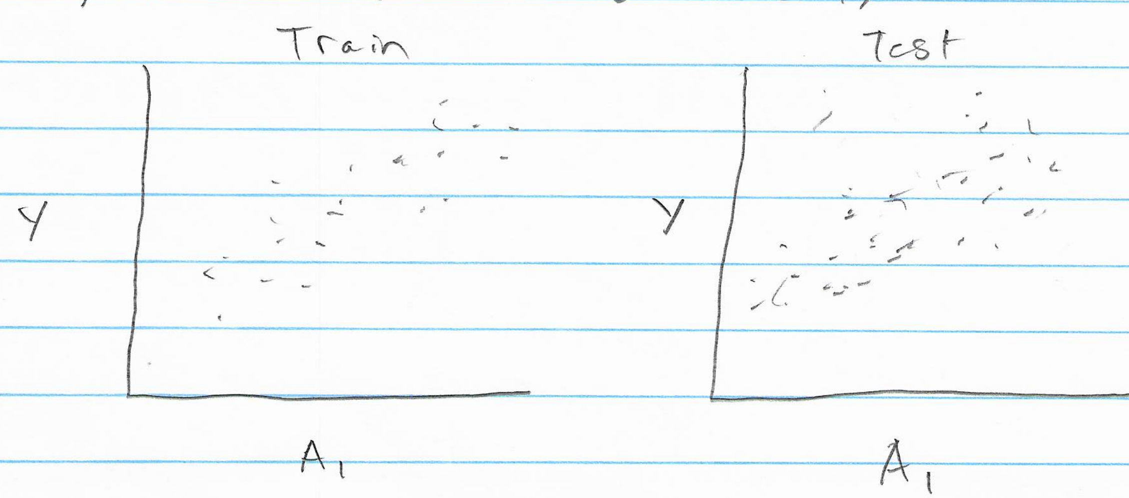
$$\hat{y} = \beta_0 + \beta_j X_{ij} \quad (\text{think } y = mx + b)$$

Process for Simple linear Regression:

1) Separate the dataset into training dataset and test dataset



2) Plot data for $j = 1$ (A_1)



3) Estimate β_0 (intercept) and β_1 (slope) on the training dataset for $j=1$ (A_1) using least squares

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_{i1} - \bar{A}_1)(y_i - \bar{y})}{\sum_{i=1}^{n} (x_{i1} - \bar{A}_1)^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{A}_1$$

where $\bar{A}_1$ and $\bar{y}$ are the means of A_1 and y, respectively

4) Calculate the sum of squared error on the test data set

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

5) Repeat for each attribute A_j in the data set ($j=2,3,\ldots m$)

6) Select the model with the lowest error

Provides a good baseline for more sophisticated linear models