TIM245 Lecture 9 (5/3/17)

Agenda

1) Evaluation of Prediction Models

2) General comments around linear regression and prediction models

3) Types of classification models

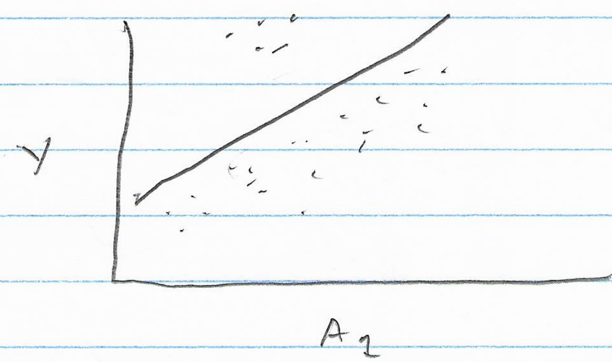
4) Logistic Regression

5) Evaluation of classification models

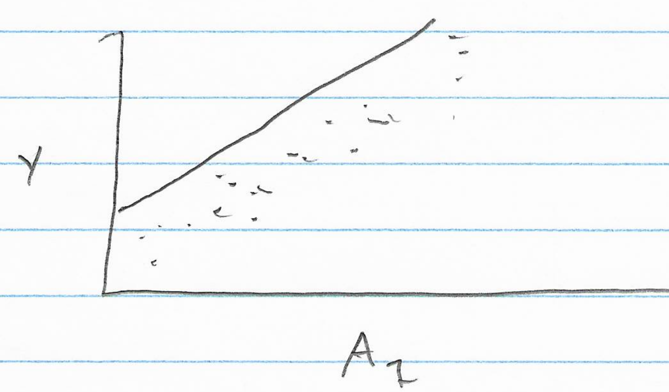
6) Return graded HW 1, Phase III of the project, HW 2

① Evaluation of Linear Models

The model training process produces the model (coefficients) that minimize $RSS(\beta)$ for the training dataset



The model testing process measure the quality of the predictions made by the model on a new unseen test data set.



Goal is to avoid selecting a model that overfits the data.

Measures of Error

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

Mean Squared Error $= \frac{1}{n} \sum_{i=1}^{n} (y - \hat{y})^2$

Mean Absolute Error (MAE) $= \frac{1}{n} \sum_{i=1}^{n} |y - \hat{y}|$

Relative Absolute Error (RAE) $= \frac{1}{n} \sum_{i=1}^{n} \frac{|y - \hat{y}|}{y}$

Correlation Coefficient $(R^2) = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$

Two Sources of error

Bias

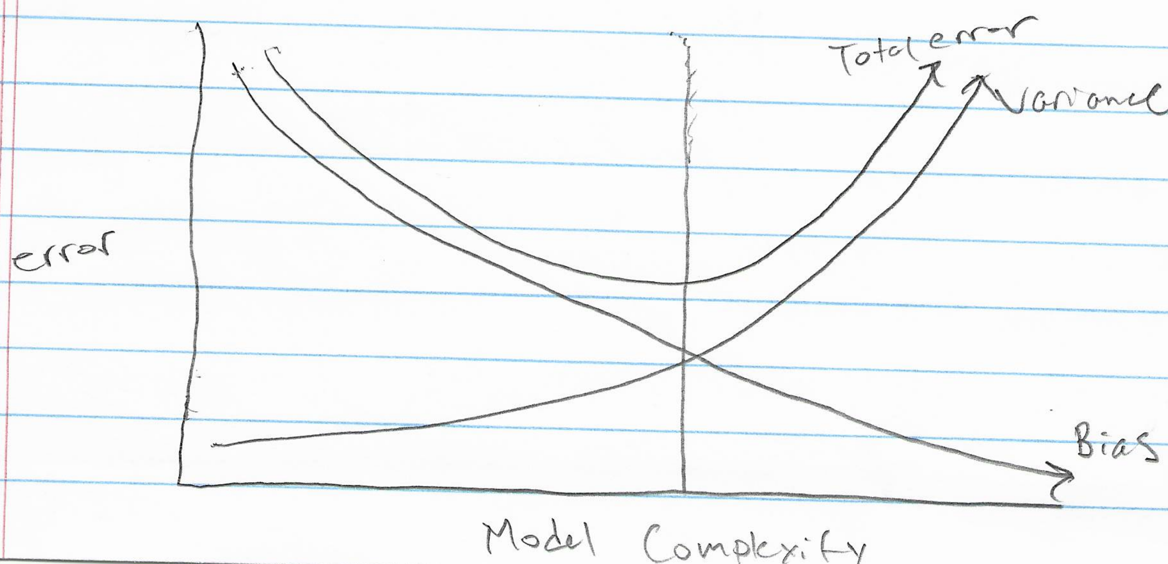
$\sum_{i=1}^{n} \hat{y} - y$

(Assumptions in algorithm)

Variance

$\sum_{i=1}^{n} (\hat{y}^2) - \left(\sum_{i=1}^{n} \hat{y}\right)^2$

(Sensitivity to fluctuations)



② General Comments on Linear Regression

1) Always plot the data and check that it is realistic given linear regression's assumptions:

   - Linear and additive (independent) relationships
   - Minimal colinearity
   - No autocorrelation (time series models)
   - Normality (Log transform can help)

2) Be careful with categorical/nominal attributes. Each value becomes a new binary attribute (dummy variable)

   e.g. $\{rain, sunny\} \rightarrow$ $rain = \{1, 0\}$
   $sunny = \{1, 0\}$

3) Two general ways of extending linear models beyond least squares

   a) Model assumptions about the errors

   e.g. $\hat{Y} = \beta_0 + \beta_1 X_{i1} + \varepsilon$

   where ε is some distribution that captures prior knowledge

b) transform the independent variables

e.g. $\hat{y} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2$

③ Types of Classification Models

Objectives: Create a model Φ that can classify an input instance $X_1, X_2, \ldots X_m$ with a label $y = \{C_1, C_2, \ldots C_s\}$

Example:

Strike	Lecture
no	yes
no	yes
yes	no
yes	yes

Two different approaches to Learning Φ

1) Generative Models: Learn the joint probability distribution $P(X, Y)$

		Lecture: yes	Lecture: no
Strike	yes	0.25	0.25
	no	0.5	0

Apply Bayes Rule to get the Conditional probability $P(Y|X)$

$$P(C_j|X) = \frac{P(X|C_j)\,P(C_j)}{\sum_{i=1}^{r} P(X|C_i)\,P(C_i)}$$

Example: Naive Bayes

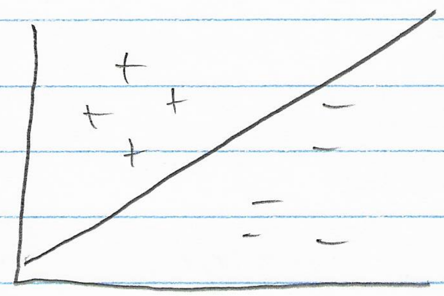
2) <u>Discriminative Models</u>: Learn the conditional probability $P(Y|X)$ directly

Lecture

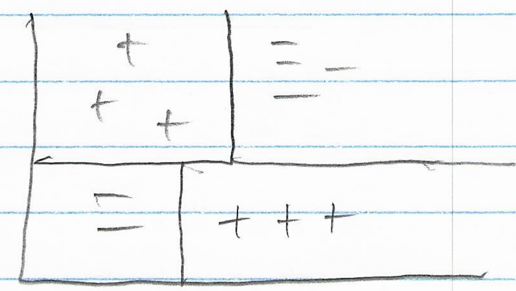
Strike		yes	no
	yes	0.5	0.5
	no	1	0

Example: Logistic Regression

Learn the decision boundry between Classes (direct map between inputs and Class label)



SVM, NN



Decision Trees

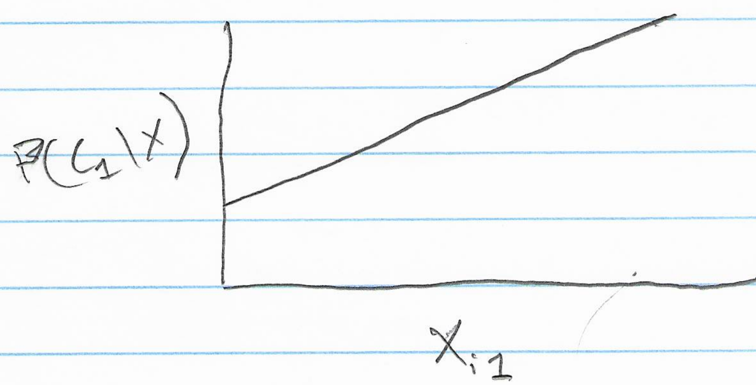
④ Logistic Regression

<u>Objective</u>: model $P(Y|X)$ as a function of X. For a binary classification problem, $Y = \{C_1, C_2\}$ we can output C_1 when $P(C_1|X) \geq 0.5$ and output C_2 when $P(C_1|X) < 0.5$

How can we model $P(C_1|X)$ as a function of X?

1) Linear relationship

$$P(C_1|X) = \beta_0 + \beta^T X$$

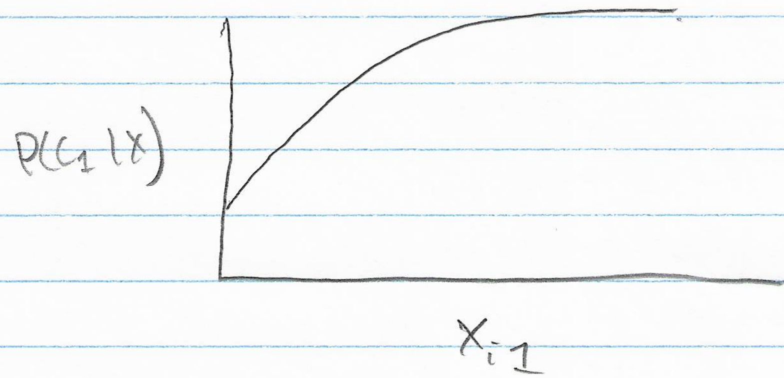


Every increment in X_{ij} either adds or subtracts to the probability

Issues: - $P(C_1|X)$ is not bounded $[0, 1]$
- Does not model diminishing returns

2) Log Function

$$P(C_1 | X) = \text{Log}(\beta_0 + \beta^T X)$$



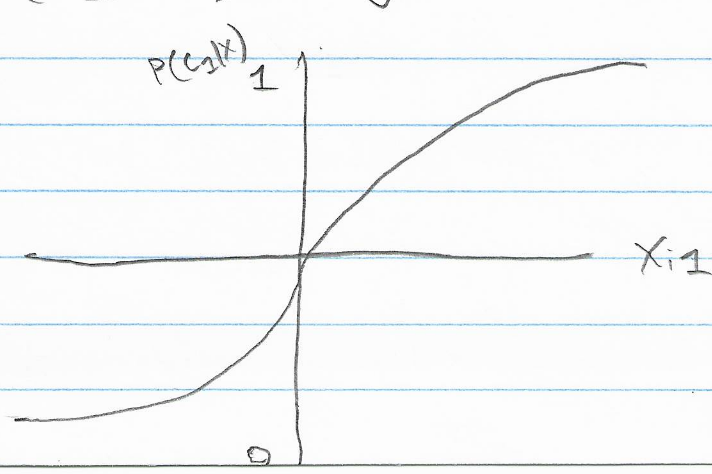
Every increment in X_{ij} multiplies the probability by a fixed amount

Issues: not unbounded in both directions

3) Logit

$$\text{Logit}(X) = \log\left(\frac{X}{1-X}\right)$$

$$P(C_1 | X) = \text{logit}(\beta_0 + \beta^T X)$$



Find the Coefficient values $\beta_0, \beta_1, \beta_2 .. \beta_m$ that have the highest likelihood given the training dataset.

⑤ Evaluation of Classification Models

Two key ideas when evaluating Classification models.

For a given C_j

<u>Precision</u> (specificity) : given a positive prediction made by the model, how likely is it to be correct

Measured as : $\dfrac{TP}{TP + TN}$

Recall (Sensitivity) : given a positive instance, how likely is it to be detected by the model

Measured as : $\dfrac{TP}{TP + FP}$

There is an inherant trade off between precision and recall

$\uparrow$ precision $\downarrow$ recall or $\downarrow$ precision $\uparrow$ recall

We need a way of balancing precision and recall

$$\text{F-Measure (F1 Score)} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Compute for each class and average the results