Evaluation and Error analysis
Validation and Regularization

Some slides taken from course materials of Andrew Ng

Karl Pearson's Coefficient of Correlation

- Karl Pearson's coefficient of correlation is a helpful statistical formula that quantifies the strength between two variables.
- This coefficient value helps in determining how strong that relationship is between the two variables.

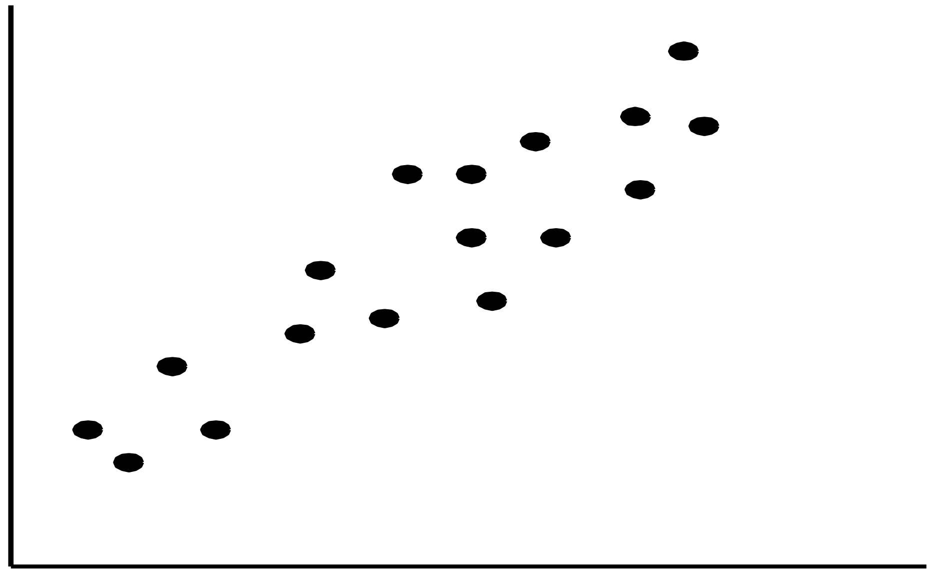
$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}}$$

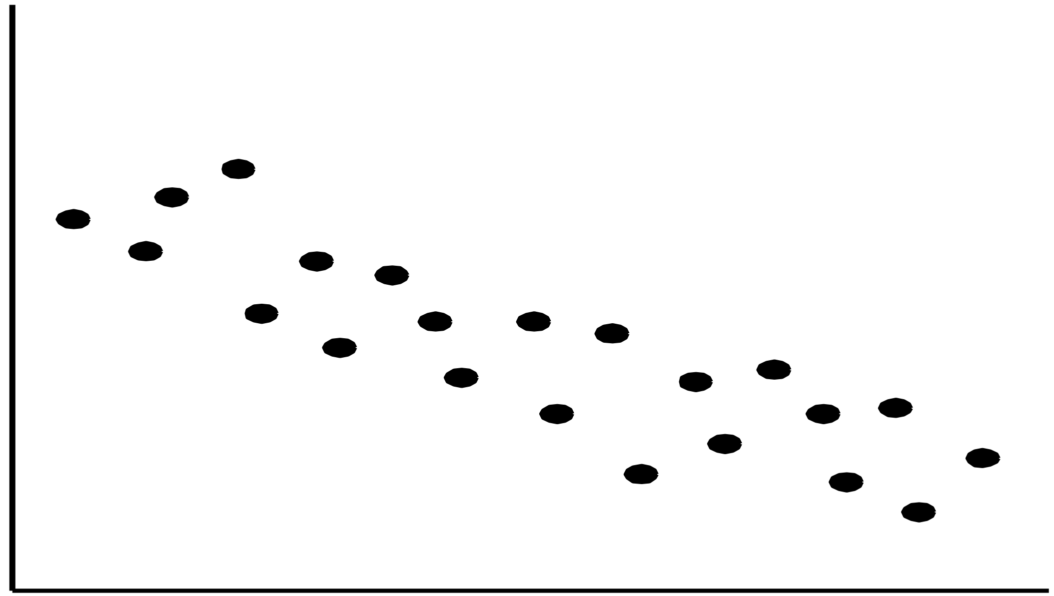
where *x* and *y* are variables and *N* is the number of instances

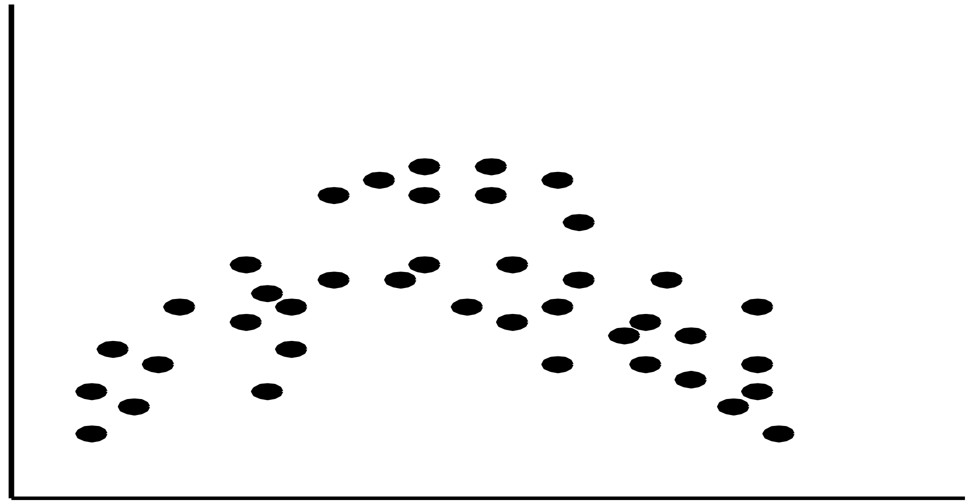
Correlation Based on Karl Pearson

- It has a value between +1 and −1
- 1 is total positive linear correlation
- 0 is no linear correlation
- −1 is total negative linear correlation

Positively and Negatively Correlated Data







- The left half fragment is positively correlated
- The right half is negative correlated

Evaluation of correlation coefficient

X	Y
95	85
85	95
80	70
70	65
60	70

How to evaluate a model?

- Regression
  - Some measure of how close are predicted values (by a model) to the actual values
- Classification
  - Whether predicted classes match the actual classes

Evaluation metrics for Regression

- Mean Squared Error (MSE)
  - For every data point, compute error (distance between predicted value and actual value)
  - Sum squares of these errors, and take average
  - More popular variant: RMSE (square root of MSE)
- R2 or R-squared
  - R-squared gives information about the goodness-of-fit measure for linear regression models
  - indicates percentage of variance in the dependent–independent variable pair.
  - Best possible R2 is 1; can be negative for a really bad model

R2 or R-squared

- Dataset has n instances $<x_i , y_i>$, i=1..N
- Predicted values: f_i, i=1..N
- Mean of actual values: $\bar{y}$

$$R^2 \equiv 1 - \frac{SS_{\rm res}}{SS_{\rm tot}}$$

$$SS_{\rm res} = \sum_i (y_i - f_i)^2$$

Residual sum of squares

$$SS_{\rm tot} = \sum_i (y_i - \bar{y})^2$$

Total sum of squares (proportional to variance)

Computing the Relationship Model

Regression model is—
Y= 0.6746*X -38.4551

X	Y	f_i	$(y_i-f_i)^2$	$(y_i-\bar{y})^2$
151	63	63.411	0.17	5.29
174	81	78.927	4.30	246.49
138	56	54.641	1.85	86.49
186	91	87.022	15.82	660.49
128	47	47.895	0.80	334.89
136	57	53.292	13.75	68.89
179	76	82.3	39.69	114.49
163	72	71.506	0.24	44.89
152	62	64.086	4.35	10.89
131	48	49.919	3.68	299.29
	$\bar{Y}$ =65.3		$\sum$ =84.65	$\sum$=1872.1

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2$$

Regression Problems

From the table below established the relationship between X and Y using linear regression analysis. Also find R squared and standard error of estimate.

X	Y
95	85
85	95
80	70
70	65
60	70

Evaluation metrics for classification

- Confusion Matrix
- Accuracy
- Precision
- Recall
- F-Score

Evaluation metrics for classification

- Let y = actual class, h = predicted class for an example
- Accuracy: Out of all examples, for what fraction is h = y?
- But accuracy is often not sufficient to indicate performance in practice

Skewed classes

- Often the class of interest is a rare class (y=1)
  - Spam emails / social network accounts
  - Cancerous cells
  - Fraud credit card transactions

Skewed classes

- Often the class of interest is a rare class (y=1)
  - Spam emails / social network accounts
  - Cancerous cells
  - Fraud credit card transactions
- Precision: Out of all examples for which model predicted h=1, for what fraction is y=1?
- Recall: Of all examples for which y=1, for what fraction did model correctly predict h=1?

Confusion Matrix

- **True Positive:**
- Interpretation: You predicted positive and it's true.
- Classifier predicted that a tumour is malignant and it actually is malignant.
- **True Negative:**
- Interpretation: You predicted negative and it's true.
- Classifier predicted that a tumour is not malignant (benign) and it actually is not malignant.
- **False Positive: (Type 1 Error)**
- Interpretation: You predicted positive and it's false.
- Classifier predicted that tumour is malignant but it actually is not.
- **False Negative: (Type 2 Error)**
- Interpretation: You predicted negative and it's false.
- Classifier predicted that tumour is not malignant (benign) but it actually is malignant.
- We describe predicted values as Positive and Negative and actual values as True and False

Model Evaluation

ID	age	income	student	credit rating	Actual	Predicted
1	youth	high	no	fair	no	yes
2	youth	high	no	excellent	no	no
3	middle aged	high	no	fair	yes	yes
4	senior	medium	no	fair	yes	yes
5	senior	low	yes	fair	yes	yes
6	senior	low	yes	excellent	no	no
7	middle aged	low	yes	excellent	yes	no
8	youth	medium	no	fair	no	no
9	youth	low	yes	fair	yes	yes
10	senior	medium	yes	fair	yes	yes
11	youth	medium	yes	excellent	yes	yes
12	middle aged	medium	no	excellent	yes	yes
13	middle aged	high	yes	fair	yes	no
14	senior	medium	no	excellent	no	yes

Predicted

Actual		Positive (Yes)	Negative (No)
	Positive (Yes)		
	Negative (No)		

Model Evaluation

ID	age	income	student	credit rating	Actual	Predicted
1	youth	high	no	fair	no	yes
2	youth	high	no	excellent	no	no
3	middle aged	high	no	fair	yes	yes
4	senior	medium	no	fair	yes	yes
5	senior	low	yes	fair	yes	yes
6	senior	low	yes	excellent	no	no
7	middle aged	low	yes	excellent	yes	no
8	youth	medium	no	fair	no	no
9	youth	low	yes	fair	yes	yes
10	senior	medium	yes	fair	yes	yes
11	youth	medium	yes	excellent	yes	yes
12	middle aged	medium	no	excellent	yes	yes
13	middle aged	high	yes	fair	yes	no
14	senior	medium	no	excellent	no	yes

		Predicted	
		Positive (Yes)	Negative (No)
Actual	Positive (Yes)	7	
	Negative (No)		

Model Evaluation

ID	age	income	student	credit rating	Actual	Predicted
1	youth	high	no	fair	no	yes
2	youth	high	no	excellent	no	no
3	middle aged	high	no	fair	yes	yes
4	senior	medium	no	fair	yes	yes
5	senior	low	yes	fair	yes	yes
6	senior	low	yes	excellent	no	no
7	middle aged	low	yes	excellent	yes	no
8	youth	medium	no	fair	no	no
9	youth	low	yes	fair	yes	yes
10	senior	medium	yes	fair	yes	yes
11	youth	medium	yes	excellent	yes	yes
12	middle aged	medium	no	excellent	yes	yes
13	middle aged	high	yes	fair	yes	no
14	senior	medium	no	excellent	no	yes

		Predicted	
		Positive (Yes)	Negative (No)
Actual	Positive (Yes)	7	2
	Negative (No)		

Model Evaluation

ID	age	income	student	credit rating	Actual	Predicted
1	youth	high	no	fair	no	yes
2	youth	high	no	excellent	no	no
3	middle aged	high	no	fair	yes	yes
4	senior	medium	no	fair	yes	yes
5	senior	low	yes	fair	yes	yes
6	senior	low	yes	excellent	no	no
7	middle aged	low	yes	excellent	yes	no
8	youth	medium	no	fair	no	no
9	youth	low	yes	fair	yes	yes
10	senior	medium	yes	fair	yes	yes
11	youth	medium	yes	excellent	yes	yes
12	middle aged	medium	no	excellent	yes	yes
13	middle aged	high	yes	fair	yes	no
14	senior	medium	no	excellent	no	yes

		Predicted	
		Positive (Yes)	Negative (No)
Actual	Positive (Yes)	7	2
	Negative (No)	2	

Model Evaluation

ID	age	income	student	credit rating	Actual	Predicted
1	youth	high	no	fair	no	yes
2	youth	high	no	excellent	no	no
3	middle aged	high	no	fair	yes	yes
4	senior	medium	no	fair	yes	yes
5	senior	low	yes	fair	yes	yes
6	senior	low	yes	excellent	no	no
7	middle aged	low	yes	excellent	yes	no
8	youth	medium	no	fair	no	no
9	youth	low	yes	fair	yes	yes
10	senior	medium	yes	fair	yes	yes
11	youth	medium	yes	excellent	yes	yes
12	middle aged	medium	no	excellent	yes	yes
13	middle aged	high	yes	fair	yes	no
14	senior	medium	no	excellent	no	yes

		Predicted	
		Positive (Yes)	Negative (No)
Actual	Positive (Yes)	7	2
	Negative (No)	2	3

Model Evaluation

- Easy to visualize via a 2 × 2 matrix (Confusion Matrix)
- Sum of diagonals = # of correct predictions
- Sum of off-diagonals = # of mistakes

		Predicted Class	
		+	-
Actual Class	+	# True Positives **TP**	# False Negatives **FN (Type-II error)**
	-	# False Positives **FP (Type-I error)**	# True Negatives **TN**

Model Evaluation

- Standard evaluation measure is classification accuracy

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

- Various other metrics are also used to evaluate classification performance

Actual Class \ Predicted Class	+	-
+	# True Positives **TP**	# False Negatives **FN (Type-II error)**
-	# False Positives **FP (Type-I error)**	# True Negatives **TN**

Model Evaluation

- Precision = Of all positive predictions, what fraction is actully positives

$$Precision\ (P) = \frac{TP}{TP + FP}$$

- Recall = Of all actual positives, what fraction is predicted as positives

$$Recall\ (R) = \frac{TP}{TP + FN}$$

		Predicted Class +	Predicted Class -
Actual Class	+	# True Positives **TP**	# False Negatives **FN** **(Type-II error)**
Actual Class	-	# False Positives **FP** **(Type-I error)**	# True Negatives **TN**

Model Evaluation

- True Positive (TP) = 7
- False Negative (FN) = 2
- False Positive (FP) = 2
- True Negative (TN) = 3
- Accuracy = ?
- Precision = ?
- Recall = ?

		Predicted	
		Positive (Yes)	Negative (No)
Actual	Positive (Yes)	7	2
	Negative (No)	2	3

Model Evaluation

- True Positive (TP) = 7
- False Negative (FN) = 2
- False Positive (FP) = 2
- True Negative (TN) = 3
- Accuracy = $\frac{7+3}{7+2+2+3} = \frac{10}{14} = 71.42\%$
- Precision = $\frac{7}{7+2} = \frac{7}{9} = 77.78\%$
- Recall = $\frac{7}{7+2} = \frac{7}{9} = 77.78\%$

		Predicted	
		Positive (Yes)	Negative (No)
Actual	Positive (Yes)	7	2
	Negative (No)	2	3

F-Score

$$\text{F-score} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

On what data to measure precision, recall, error rate, ..?

- Option 1: training set
- Option 2: some other set of examples that was unknown at the time of training (test set)

- Motivation for ML: learn a model that performs well (generalizes well) to unknown examples
- Option 2 gives better guarantees for generalization of a learnt model