

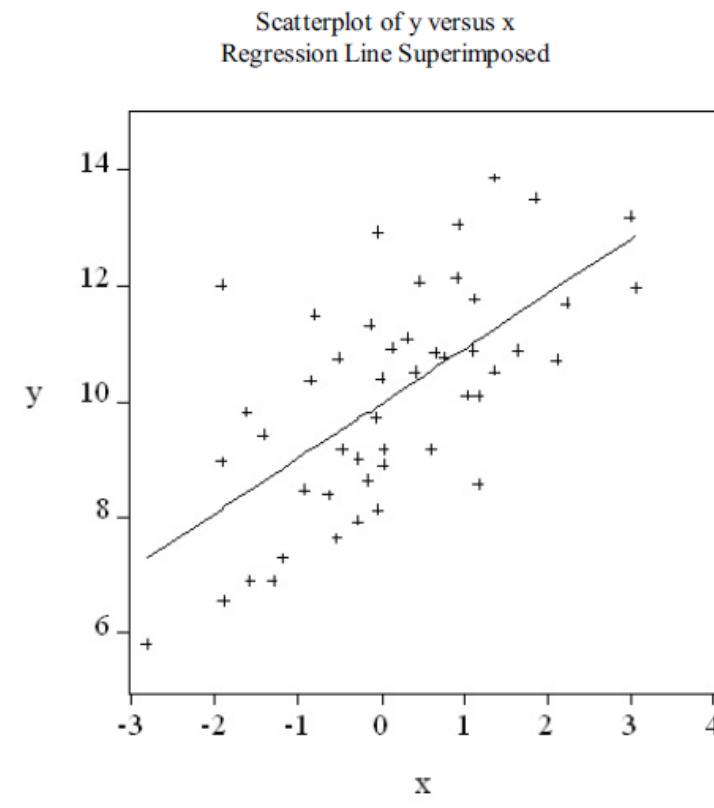
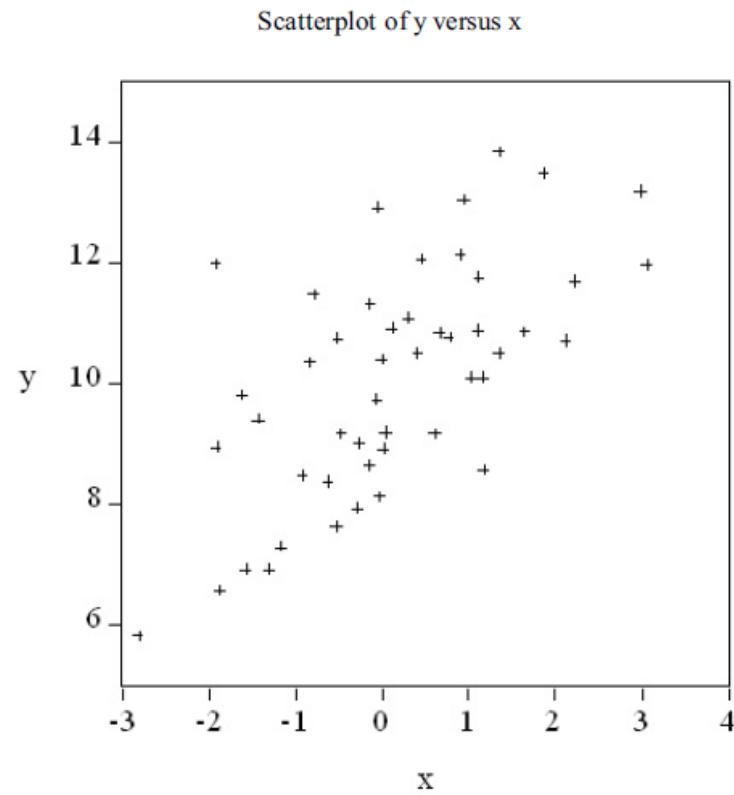
Regression Analysis

DSE 220

Regression Analysis

- One of the basic tools for forecasting
- A statistical technique to describe relationships among variables
- Consider two variables y and x
 - Describe y using x
 - y : dependent variable
 - x : independent variable (explanatory, exogenous)

Regression Analysis



Regression Analysis

- How to find the line that fits best?
 - Line: $y = \beta_0 + \beta_1 X$
 - How to find β_0 and β_1 ?
- Example

Regression Analysis

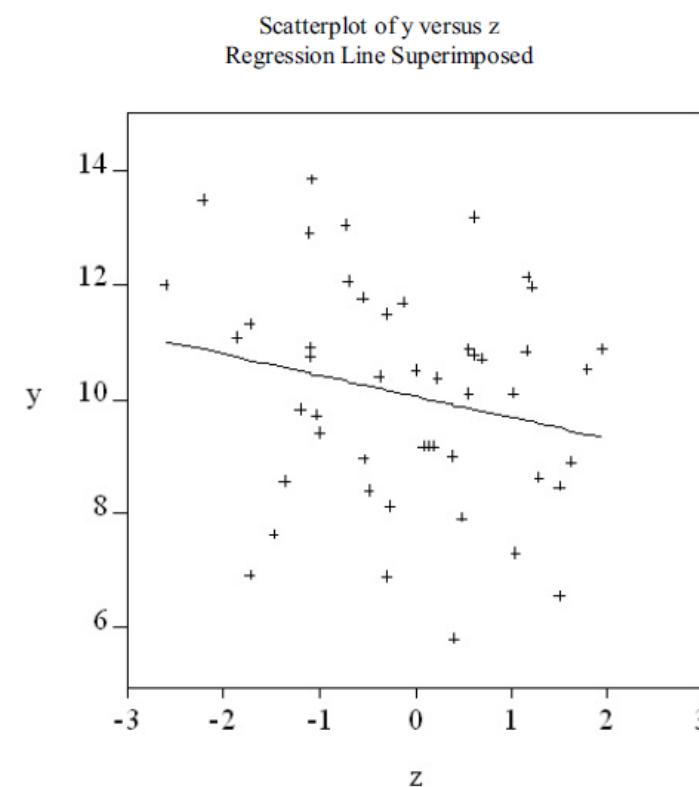
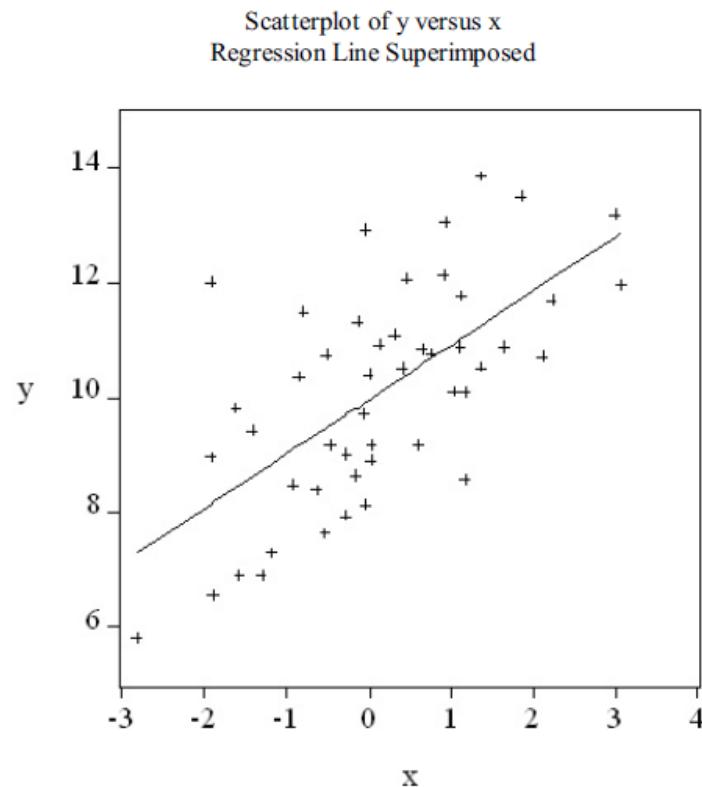
- Probabilistic Model
 - $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$
 - $\varepsilon_t \stackrel{iid}{\rightarrow} N(0, \sigma^2)$
 - Model parameters: $\beta_0, \beta_1, \sigma^2$
- If this model is correct:
 - Expected value of y conditional on $x = x^*$
 - $E(y|x^*) = \beta_0 + \beta_1 x^*$

Regression Analysis

- Fitted values
 - $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$
- Minimize residuals
 - Residuals = in-sample forecast errors
 - $e_t = y_t - \hat{y}_t$
- Least-squares estimation
 - $\sum_{t=1}^T (y_t - \hat{y}_t)^2$

Regression Analysis

- Multiple linear regression
 - y is described by more than one explanatory variable



Regression Analysis

- Multiple linear regression model

- $y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + \varepsilon_t$
- $\varepsilon_t \stackrel{iid}{\rightarrow} N(0, \sigma^2)$

- Fitted values

- $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t + \hat{\beta}_2 z_t$

- Residuals

- $e_t = y_t - \hat{y}_t$

- Least-squares estimation

- $\sum_{t=1}^T (y_t - \hat{y}_t)^2$

Goodness-of-fit Statistics

- Sum squared resid.
 - Objective of the least-squares estimation
 - Sum of squared residuals
 - Not of much value in isolation
 - Input to other diagnostics
 - Useful for comparing models and testing hypotheses

$$SSR = \sum_{t=1}^T e_t^2$$

Goodness-of-fit Statistics

- R-squared (R^2)
 - Indicates how much of $\text{var}(y)$ can be explained by the variables included in the regression
 - Intuition: $\frac{\text{var}(y|x)}{\text{var}(y)}$
 - Measurement of in-sample success of the regression equation
 - If intercept is included, $0 < R^2 < 1$

$$R^2 = 1 - \frac{\frac{1}{T} \sum_{t=1}^T e_t^2}{\frac{1}{T} \sum_{t=1}^T (y_t - \bar{y}_t)^2}$$

Goodness-of-fit Statistics

- Adjusted R-squared (\bar{R}^2)
 - Same interpretation as R^2 but formula is slightly different
 - Adjusted for degrees of freedom used in fitting the model
 - Adjustment: penalize the amount of right-hand-side variables

$$\bar{R}^2 = 1 - \frac{\frac{1}{T-k} \sum_{t=1}^T e_t^2}{\frac{1}{T-1} \sum_{t=1}^T (y_t - \bar{y}_t)^2}$$

Goodness-of-fit Statistics

- Akaike info criterion (AIC)
 - Estimate of the out-of-sample forecast error variance
 - Similar to s^2 but penalizes degrees of freedom more harshly
 - Used to compare forecasting models

$$AIC = e^{\left(\frac{2k}{T}\right)} \frac{\sum_{t=1}^T e_t^2}{T}$$

Goodness-of-fit Statistics

- Schwarz criterion (SIC)
 - Alternative to AIC
 - Harsher penalty for degrees-of-freedom
 - Used to compare forecasting models

$$SIC = T^{\left(\frac{k}{T}\right)} \frac{\sum_{t=1}^T e_t^2}{T}$$

Goodness-of-fit Statistics

- Durbin-Watson stat. (DW)
 - Errors from a good forecasting model should be unforecastable
 - Forecastable error -> room for improvement in the model
 - Correlation among errors -> forecastable information
 - DW tests if the regression disturbances over time are serially correlated.
 - $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$, where $\varepsilon_t = \varphi \varepsilon_{t-1} + v_t$
 $v_t \stackrel{iid}{\rightarrow} N(0, \sigma^2)$
 - ε_t is serially correlated when $\varphi \neq 0$.
 - Ideal case $\varphi = 0$

Goodness-of-fit Statistics

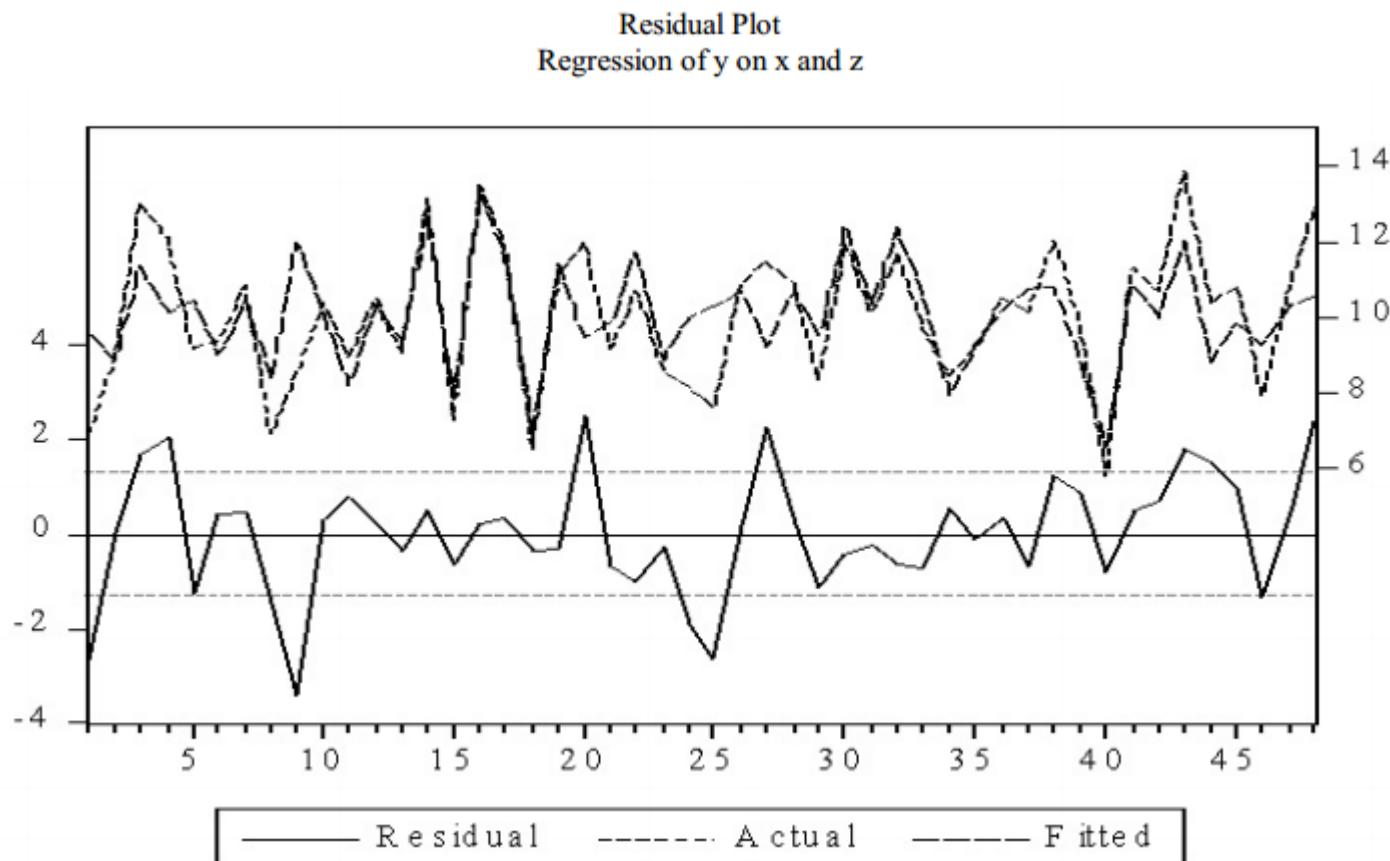
- Durbin-Watson stat. (DW)
 - $H_0: \varphi = 0$
 - $0 \leq DW \leq 4$
 - **OK:** $DW \sim 2$
 - **Alarm:** $DW < 1.5$
 - Consult DW tables for significance level for rejecting the null hypothesis

$$DW = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2}$$

Goodness-of-fit Statistics

- Residual plot
 - Examine :
 - Actual data (y_t)
 - Fitted values (\hat{y}_t)
 - Residuals (e_t)

Goodness-of-fit Statistics



Evaluation Metrics

Evaluation Metrics

- Up to now: measure a model's performance by some simple metric
 - classifier error rate, accuracy, ...
- Simple example: accuracy

$$\text{accuracy} = \frac{\text{Number of correct decisions made}}{\text{Total number of decisions made}}$$

- Classification accuracy is popular, but usually **too simplistic** for applications of data mining to real business problems
- **Decompose** and count the different types of correct and incorrect decisions made by a classifier

Unequal costs and benefits

- How much do we care about the different **errors** and correct decisions?
 - Classification accuracy makes no distinction between **false positive** and **false negative** errors
 - In real-world applications, different kinds of errors lead to different consequences!
- Examples for medical diagnosis:
 - a patient has cancer (although he does not)
→ **false positive error**, expensive, but not life threatening
 - a patient has cancer, but she is told that she has not
→ **false negative error**, more serious
- Errors should be counted separately
 - Estimate cost or benefit of each decision

Confusion Matrix

- A **confusion matrix** for a problem involving n classes
 - is an $n \times n$ matrix with the columns labeled with actual classes and the rows labeled with predicted classes

		Predicted Classes	
		p	n
True Values	1	True Positives (TP)	False Negative (FN)
	0	False Positives (FP)	True Negatives (TN)

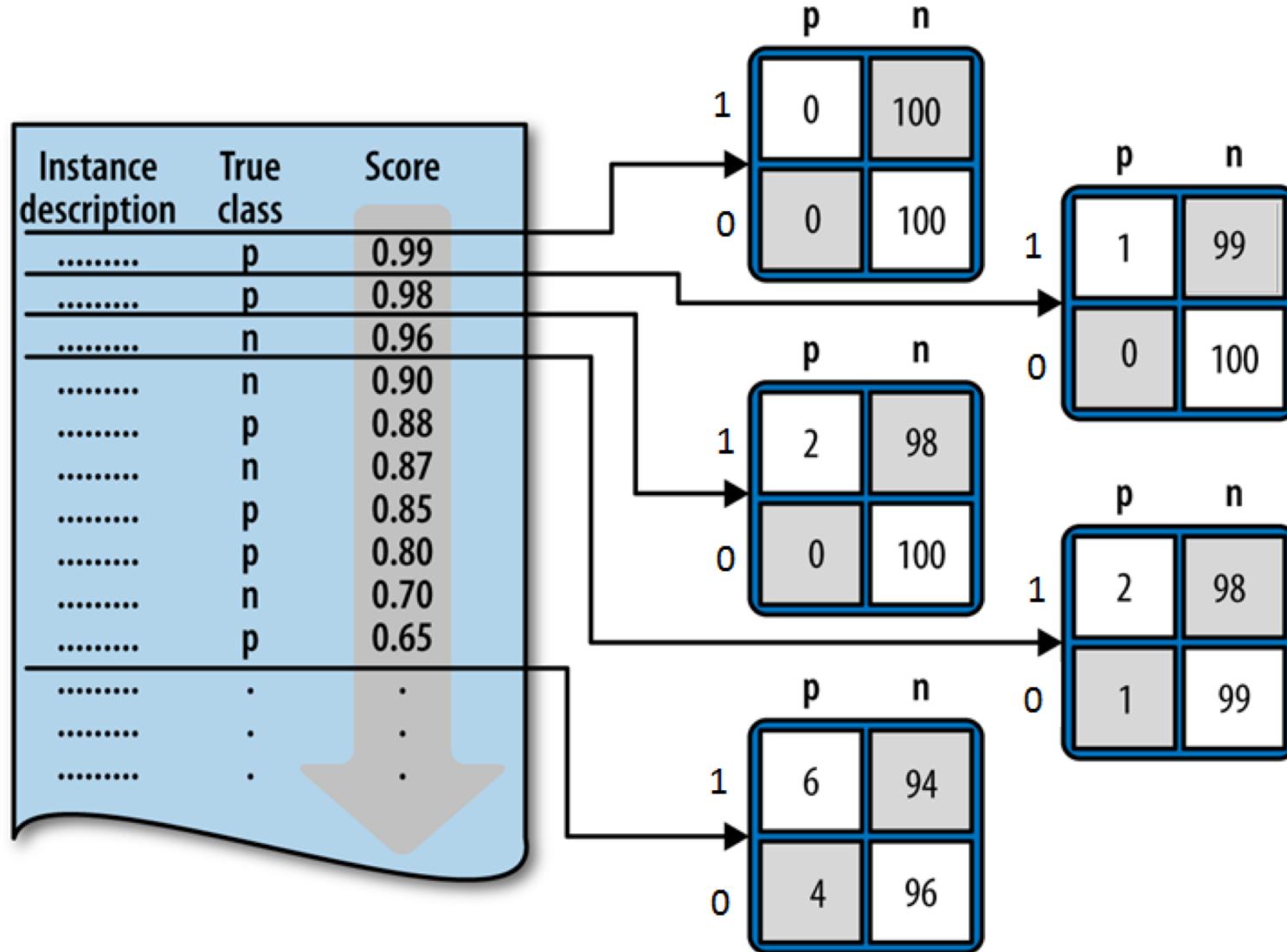
- Each example in a test set has an **actual class label** and the **class predicted** by the classifier
- The confusion matrix separates out the decisions made by the classifier
 - actual/true classes: **1** (Positive Label), **0** (Negative Label)
 - predicted classes: **p**(ositive), **n**(egative)
 - The main diagonal contains the count of correct decisions

Other Evaluation Metrics

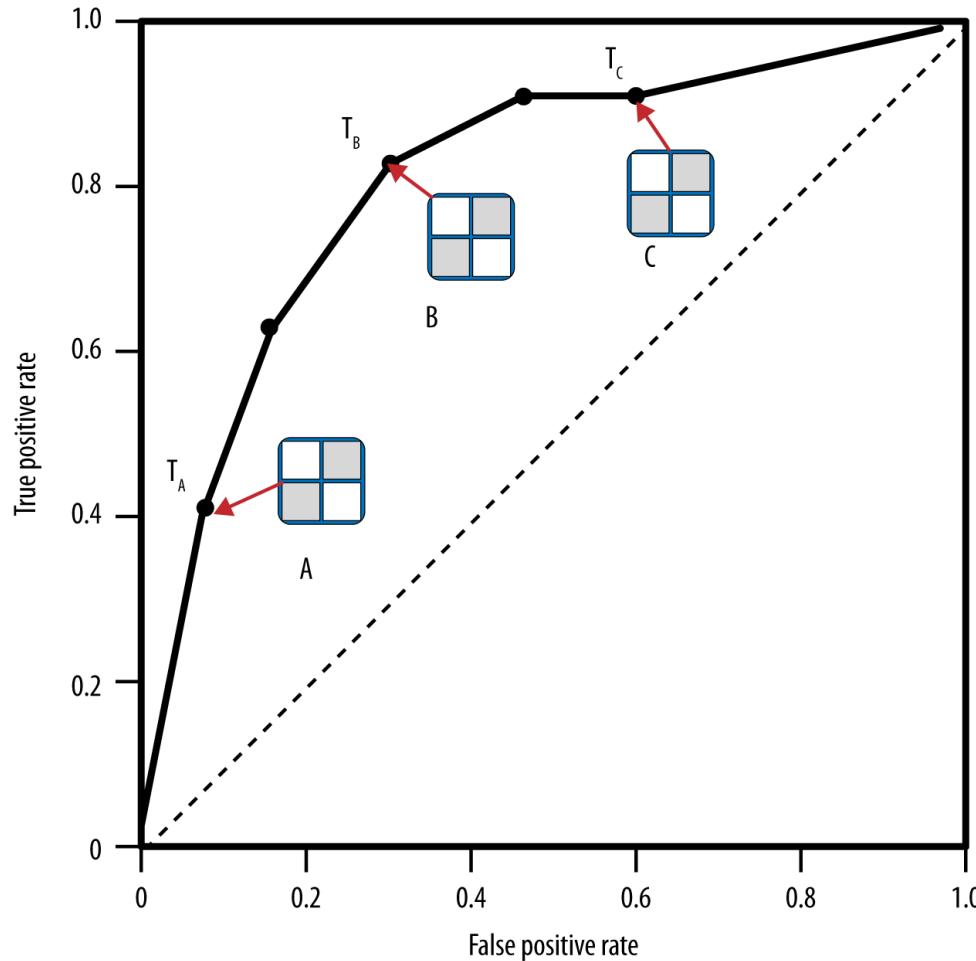
- Based on the entries of the confusion matrix, we can describe various evaluation metrics
 - True positive rate (Recall): $\frac{TP}{TP+FN}$
 - False negative rate: $\frac{FN}{TP+FN}$
 - Precision (accuracy over the cases predicted to be positive): $\frac{TP}{TP+FP}$
 - F-measure (harmonic mean): $2 \cdot \frac{precision \cdot recall}{precision + recall}$
 - Specificity: $\frac{TN}{TN+FP}$
 - Sensitivity: $\frac{TP}{TP+FN}$
 - Accuracy (count of correct decisions): $\frac{TP+TN}{P+N}$

ROC Curves

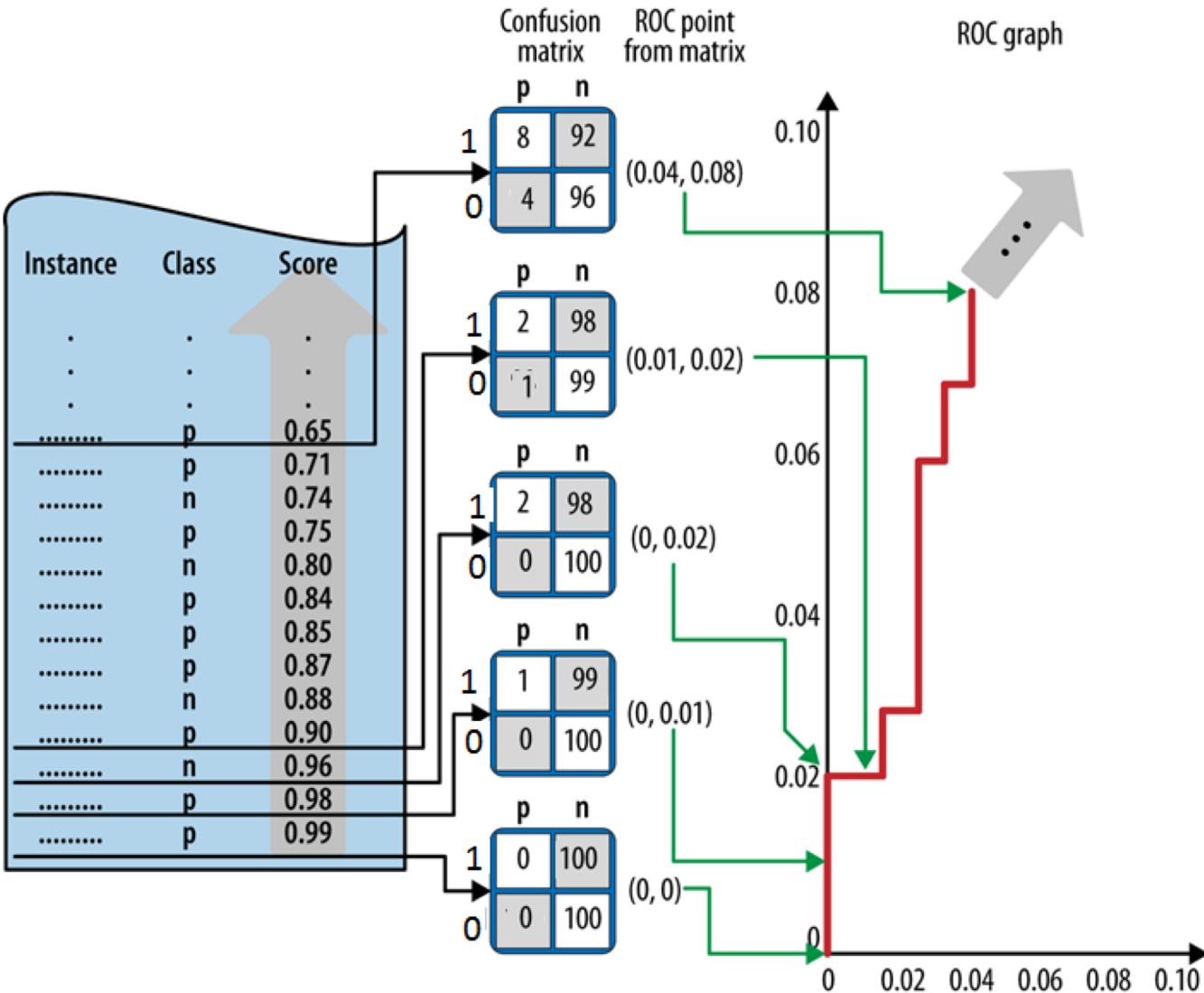
Ranking Instead of Classifying



ROC Graphs and Curves



ROC Graphs and Curves



Generating ROC curve: Algorithm

- Sort the test set by the model predictions
- Start with cutoff = max (prediction)
- Decrease cutoff, after each step count the number of true positives TP (positives with prediction above the cutoff) and false positives FP (negatives above the cutoff)
- Calculate TP rate (TP/P) and FP (FP/N) rate
- Plot current number of TP/P as a function of current FP/N

Area Under the ROC Curve (AUC)

- The area under a classifier's curve expressed as a fraction of the unit square
 - Its value ranges from zero to one
- The AUC is useful when a single number is needed to summarize performance, or when nothing is known about the operating conditions
 - A ROC curve provides more information than its area
- Equivalent to the **Mann-Whitney-Wilcoxon measure**
 - Also equivalent to the Gini Coefficient (with a minor algebraic transformation)
 - Both are equivalent to the probability that a randomly chosen positive instance will be ranked ahead of a randomly chosen negative instance

Performance Evaluation

- Training Set:

Model	Accuracy
Classification Tree	95%
Logistic Regression	93%
k -Nearest Neighbors	100%
Naive Bayes	76%

- Test Set:

Model	Accuracy	AUC
Classification Tree	91.8% \pm 0.0	0.614 \pm 0.014
Logistic Regression	93.0% \pm 0.1	0.574 \pm 0.023
k -Nearest Neighbors	93.0% \pm 0.0	0.537 \pm 0.015
Naive Bayes	76.5% \pm 0.6	0.632 \pm 0.019

Performance Evaluation

- Naive Bayes confusion matrix:

	p	n
1	127 (3%)	200 (4%)
0	848 (18%)	3518 (75%)

- k -Nearest Neighbors confusion matrix:

	p	n
1	3 (0%)	324 (7%)
0	15 (0%)	4351 (93%)

ROC Curve

