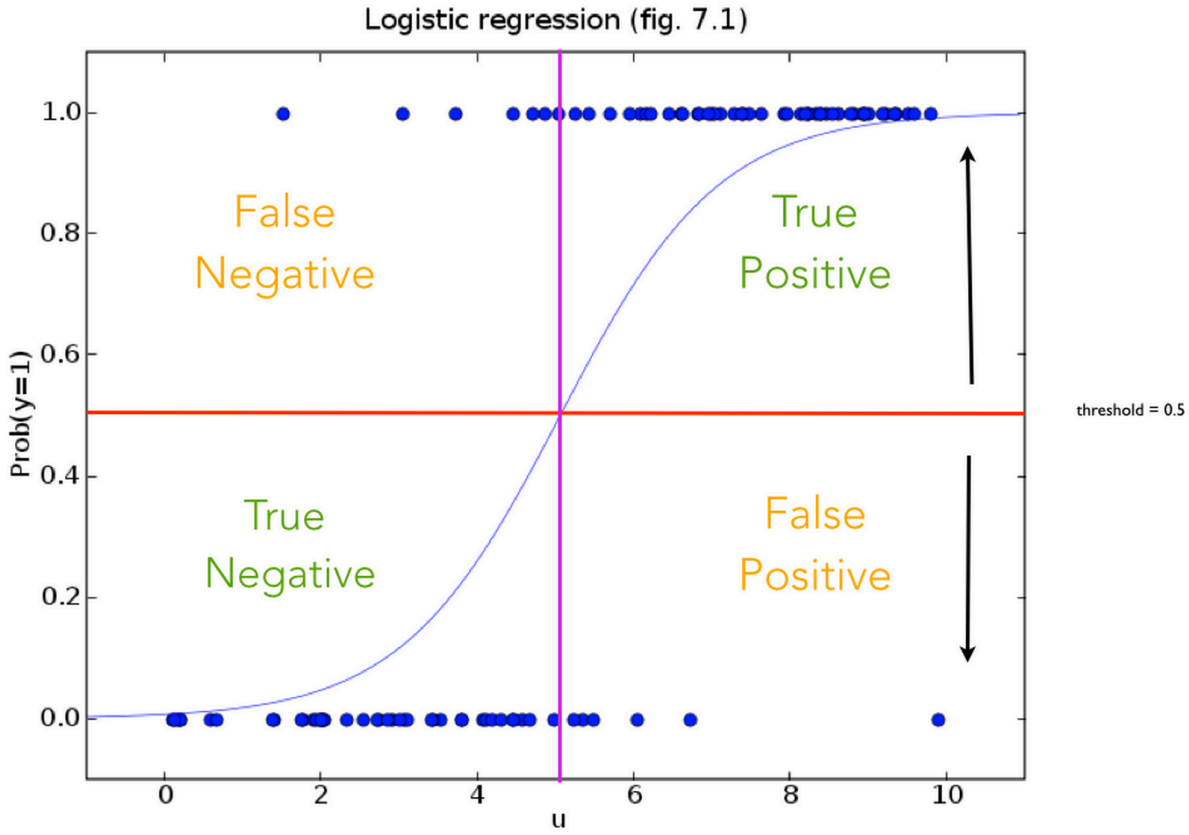
Profit Curves

Overview

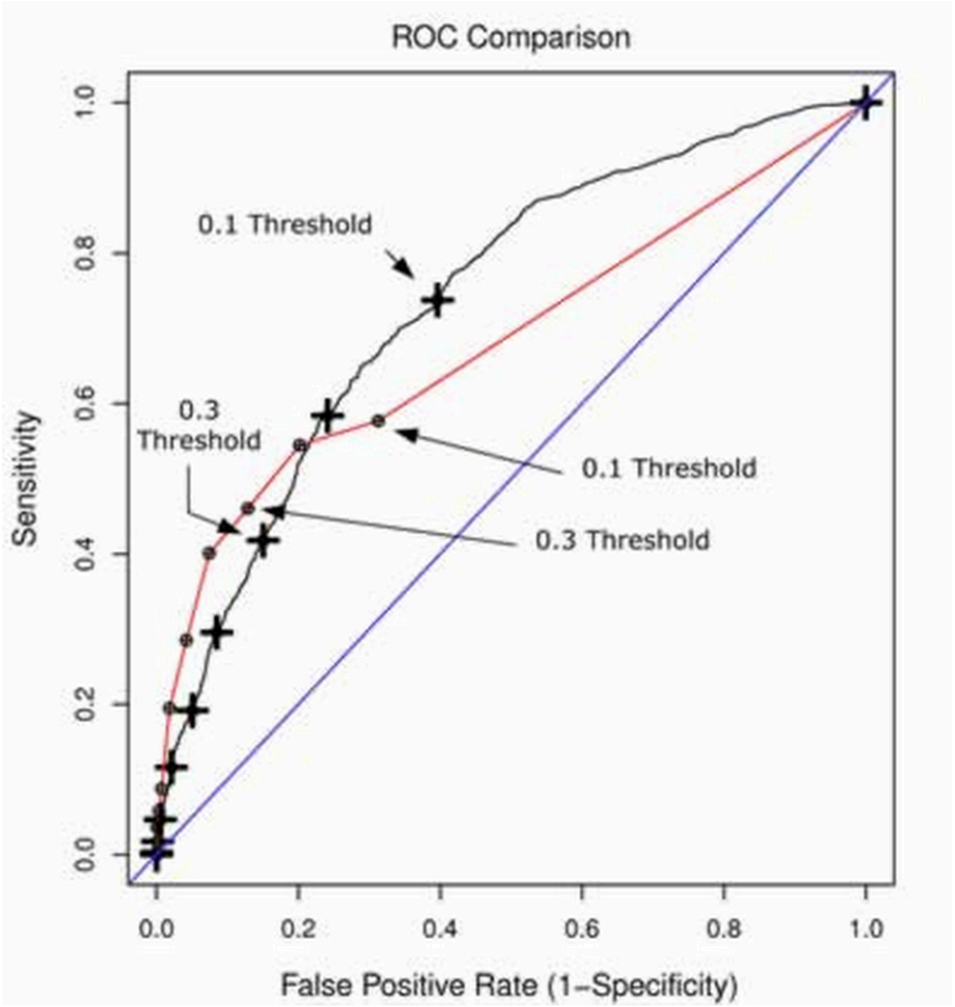
- Review of Logistic Regression and ROC Curves
- Profit Calculation
  - Simple
  - Expanded
- Profit Curves

Logistic Regression



	Predicted Yes	**Predicted No**
Actual Yes	True positive	False negative
Actual No	False positive	True negative

Logistic Regression - Evaluation



- Area under the Curve (AUC)
- F1 Score
- Precision / Recall
- Sensitivity / Specificity

Evaluation of Classifier

true positive (TP)

eqv. with hit

true negative (TN)

eqv. with correct rejection

false positive (FP)

eqv. with false alarm, Type I error

false negative (FN)

eqv. with miss, Type II error

accuracy (ACC)

$$ACC = (TP + TN)/(P + N)$$

F1 score

is the harmonic mean of precision and sensitivity

$$F1 = 2TP/(2TP + FP + FN)$$

sensitivity or true positive rate (TPR)

eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

specificity (SPC) or true negative rate (TNR)

$$SPC = TN/N = TN/(FP + TN)$$

precision or positive predictive value (PPV)

$$PPV = TP/(TP + FP)$$

negative predictive value (NPV)

$$NPV = TN/(TN + FN)$$

fall-out or false positive rate (FPR)

$$FPR = FP/N = FP/(FP + TN)$$

false discovery rate (FDR)

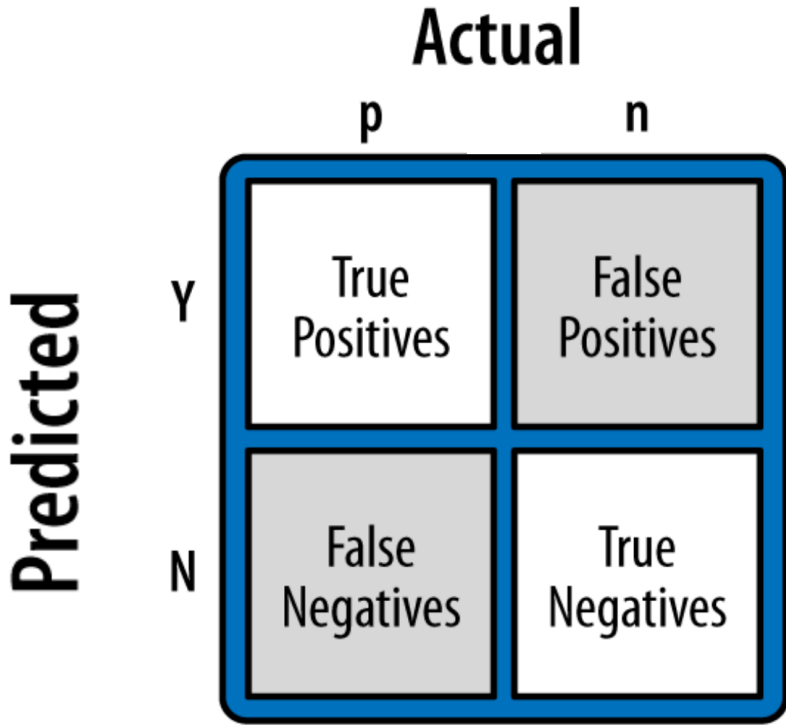
$$FDR = FP/(FP + TP) = 1 - PPV$$

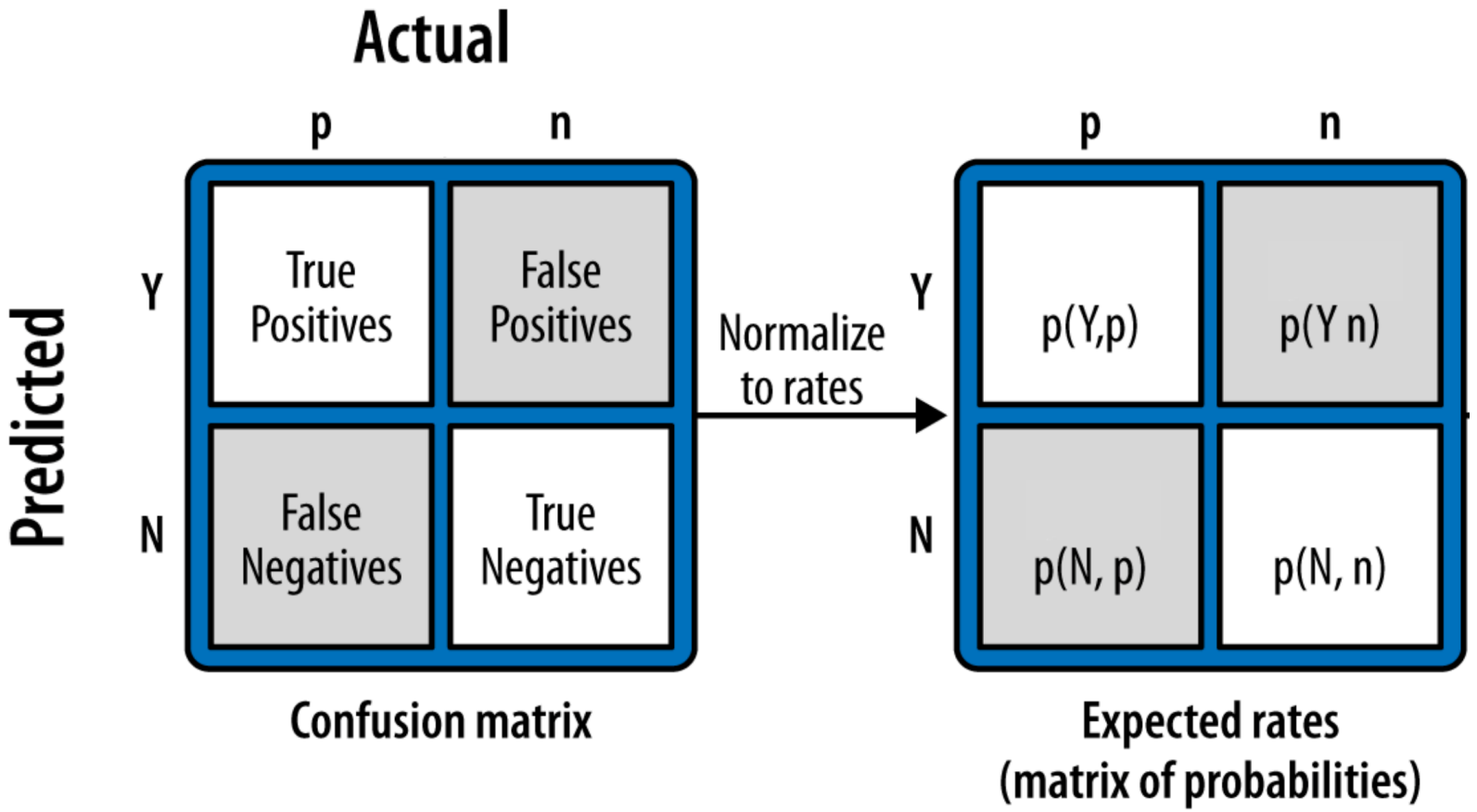
false negative rate (FNR)

$$FNR = FN/(FN + TP) = 1 - TPR$$

Confusion Matrix

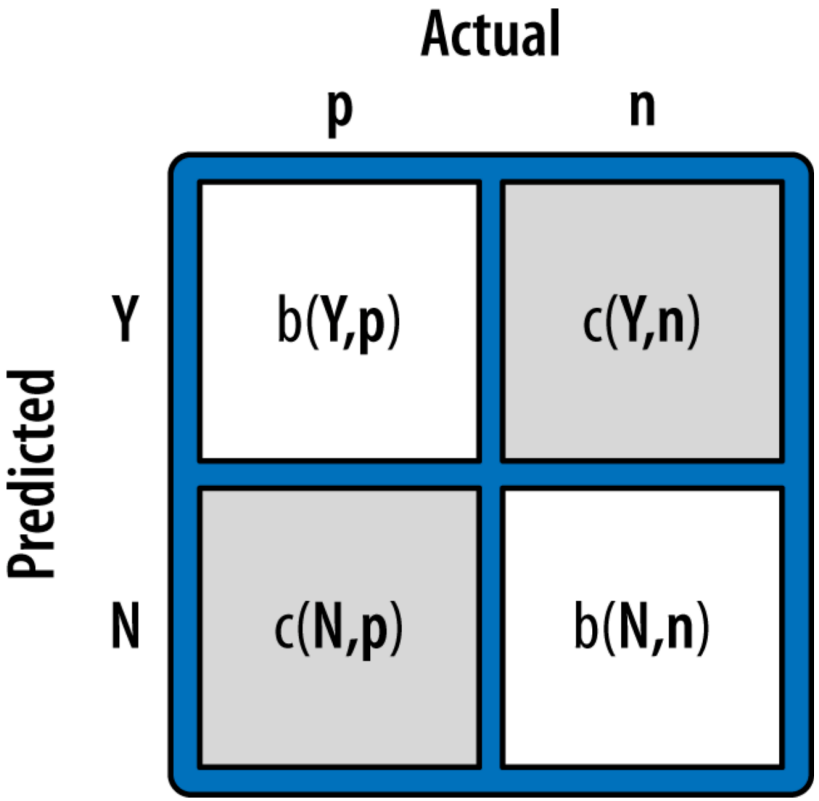
- ROC Curves assume equal cost to misclassification
  - However different errors have different costs
  - Correct classification can have different benefits




Actual
p
n
Predicted
Y
N
True Positives
False Positives
False Negatives
True Negatives
Confusion matrix
Normalize to rates
p
n
Y
N
p(Y,p)
p(Y n)
p(N, p)
p(N, n)
Expected rates (matrix of probabilities)

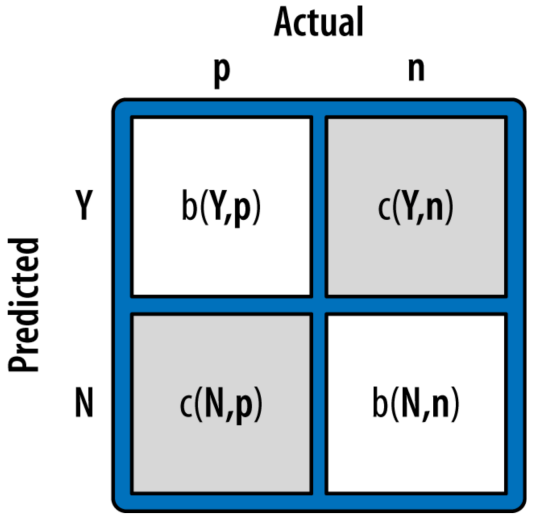
Cost benefit Matrix

- Defined from the business situation:



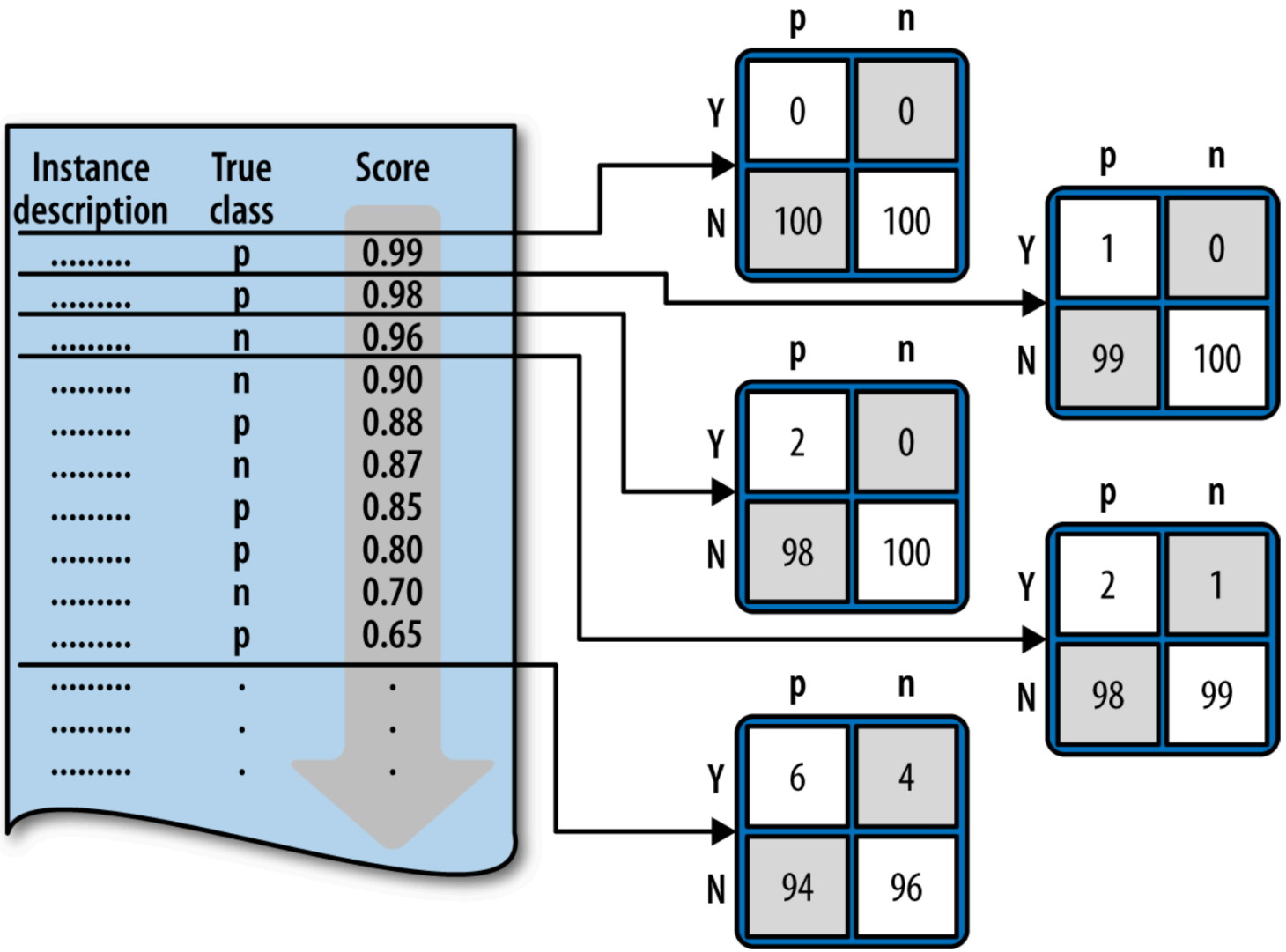
Profit Calculation

Combining information from the Confusion matrix and the Cost-Benefit matrix we can calculate Expected Profit!



$$
\begin{aligned}
E[Profit] &= P(Y,p)\cdot b(Y,p) + P(Y,n)\cdot c(Y,n) + \\
&\quad P(N,p)\cdot c(N,p) + P(N,n)\cdot b(N,n) \\
&= P(Y|p)\cdot P(p)\cdot b(Y,p) + P(Y|n)\cdot P(n)\cdot c(Y,n) + \\
&\quad P(N|p)\cdot P(p)\cdot c(N,p) + P(N|n)\cdot P(n)\cdot b(N,n) \\
&= P(p)\cdot [P(Y|p)\cdot b(Y,p) + P(N|p)\cdot c(N,p)] + \\
&\quad P(n)\cdot [P(Y|n)\cdot c(Y,n) + P(N|n)\cdot b(N,n)]
\end{aligned}
$$

Building the Profit Curve



Building the Profit Curve

For a given model f, each threshold value T gives a point on the Profit Curve

Model score is the threshold probability classifying $+$ vs $-$

1. Allow T to be the maximum score
2. $TP = 0, FP = 0$
3. Calculate $E[Profit]$
4. For each observation, i:
   - If $\hat{\pi}_i > T \longrightarrow$ increment TP
   - Else $\longrightarrow$ increment FP
5. Add point (% Test Instances predicted Positive, $E[Profit]$) to the Profit Graph

Increment T from max-score to min-score, repeating steps 1-4

Profit Curves!