Logistic Regression

Topics

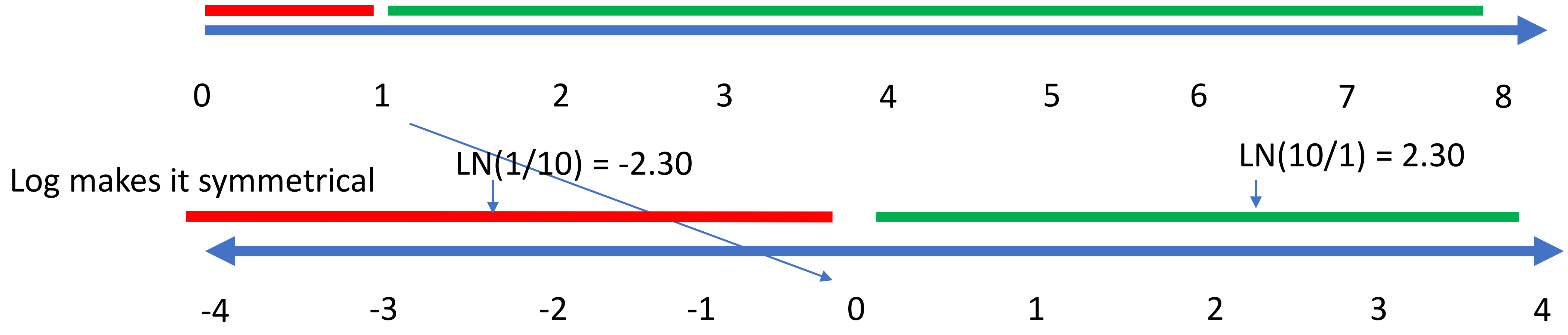
- Odds
- Binary Dependent Variables
- Logistic Regression
- Prediction and Confusion Matrix
- Sensitivity, Specificity, and ROC-AUC Curve

Odds

- A way to express Likelihood that an event will take place
- Can be written as X to Y or X:Y or X/Y
- Odd For:
  - A ratio of probabilities: Odds For = (probability that the event will happen)/(probability that the event will not happen)
  - Odds(Y=1) = P/(1-P)
- Usually written as Example 3:1 odds : 3/1 = P/(1-P) -> 3(1-P)=P -> 3-3P = P -> 3=4P
- P=3/4 or 75%
- Or P = (odds for) /(1+odds for)

Odds vs Log Odds

- Prob event will happen/Prob event will not happen
- 0.1/0.9 = 0.111 , 0.2/0.8=0.25, 0.3/0.7 = 0.429
- 0.5/0.5 = 1
- 0.7/0.3 = 2.33 , 0.8/0.2 = 4, 0.9/0.1 = 9



Binary Dependent Variables

- Yes/No
- True/False

Examples:

Medical: Pregnant or not, Cancerous or not

Transportation: Cancel flight or not

Purchasing: Award contract or not

Fraud Detection: Legitimate credit card transaction or not

Logistic Regression

- Similar to Linear Regression
- Response variable is binary
- Results are expressed as a probability between 0 and 1

- Multinomial Logistic Regression (sometimes called multi-class classification) using a one vs all (softmax)....beyond the scope of this course

Logistic Regression

- Using the coefficients and intercept we create **the logistic function**; we can calculate the maximum likelihood to fit the model:

$$p(X) = \frac{e^{\beta_0+\beta_1*pred_1+\ldots+\beta_8*pred_8}}{1+e^{\beta_0+\beta_1*pred_1+\ldots+\beta_8*pred_8}}$$

- We then can find the odds:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1*pred_1+\ldots+\beta_8*pred_8}$$

- Finally, for the simplified model, we can calculate the log-odds (logit):

$$log(\frac{p(X)}{1-p(X)}) = \beta_0+\beta_1*pred_1+\ldots+\beta_8*pred_8$$

Probability to Log Odds

- The transformation maps probability form 0 and 1 to log odds ranging from negative infinity to positive infinity

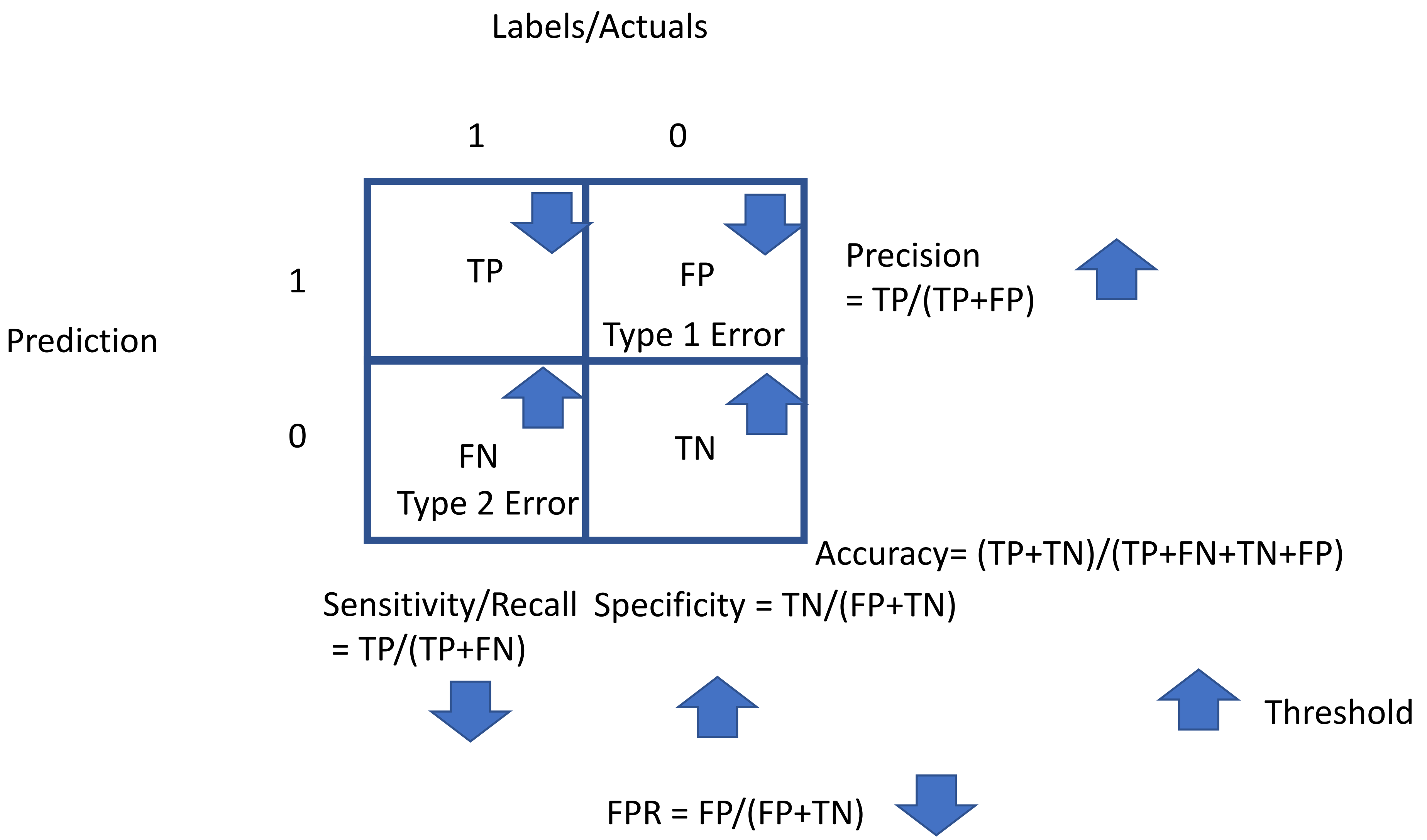
Log-odds	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
Probability	0.002	0.007	0.018	0.047	0.119	0.269	0.500	0.731	0.881	0.953	0.982	0.993	0.998

Interpreting the Logistic Regression Model

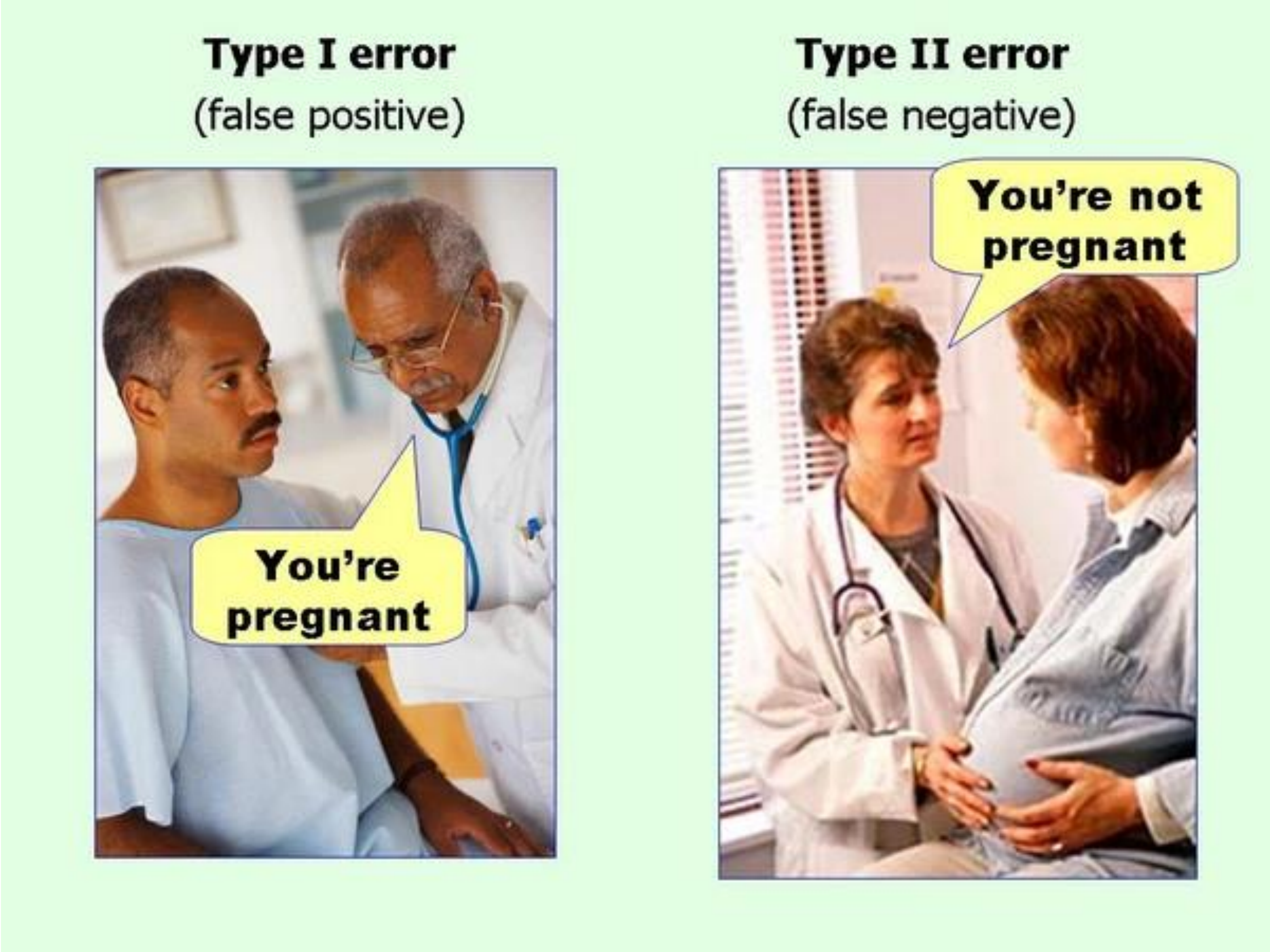
$$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 * pred_1$$

- One unit increase in X, the natural log odds will increase by coefficient B1
- The exact change in odds is (exp(B1)-1)*100%

Confusion Matrix



Type I vs Type II Error



Example Of Changing Thresholds

Thresh	TP	TN	FP	FN	Recall	Percision	Specificity	FPR	Accuracy
1	0	183	0	317	0.0%	100.0%	100.0%	0.0%	36.6%
0.9	237	183	0	80	74.8%	100.0%	100.0%	0.0%	84.0%
0.8	237	183	0	80	74.8%	100.0%	100.0%	0.0%	84.0%
0.7	240	183	0	77	75.7%	100.0%	100.0%	0.0%	84.6%
0.6	255	175	8	62	80.4%	97.0%	95.6%	4.4%	86.0%
0.5	280	157	26	37	88.3%	91.5%	85.8%	14.2%	87.4%
0.4	290	138	45	27	91.5%	86.6%	75.4%	24.6%	85.6%
0.3	298	124	59	19	94.0%	83.5%	67.8%	32.2%	84.4%
0.2	310	88	95	7	97.8%	76.5%	48.1%	51.9%	79.6%
0.1	317	33	150	0	100.0%	67.9%	18.0%	82.0%	70.0%
0	317	0	183	0	100.0%	63.4%	0.0%	100.0%	63.4%

ROC-AUC

- Receiver operating characteristic curve is a graph that summarizes the performance of a binary classification model on the positive class. The x-axis indicates the 1-Specificity (or false positive rate) and the y-axis indicates Recall or the true positive rate.
- Ideally, we would want TPR =1 and FPR =0
- AUC can be calculated to give a single score across all threshold values
- How to interpret the AUC?
  - It is the probability that the scores given by the model will rank a randomly chosen positive instance higher than the randomly chosen negative instance
- Useful in comparing different models

ROC-AUC Example

Thresh	TP	TN	FP	FN	Recall	FPR
1	0	183	0	317	0	0
0.9	237	183	0	80	0.747634	0
0.8	237	183	0	80	0.747634	0
0.7	240	183	0	77	0.757098	0
0.6	255	175	8	62	0.804416	0.043715847
0.5	280	157	26	37	0.883281	0.142076503
0.4	290	138	45	27	0.914826	0.245901639
0.3	298	124	59	19	0.940063	0.322404372
0.2	310	88	95	7	0.977918	0.519125683
0.1	317	33	150	0	1	0.819672131
0	317	0	183	0	1	1