Machine Learning: Yoshinari Fujinuma

University of Colorado Boulder

LECTURE 12

Slides adapted from Chenhao Tan, Chris Ketelsen

Logistics

- HW2 due this Friday
- No class on Wednesday (Wellness day)

Learning objectives

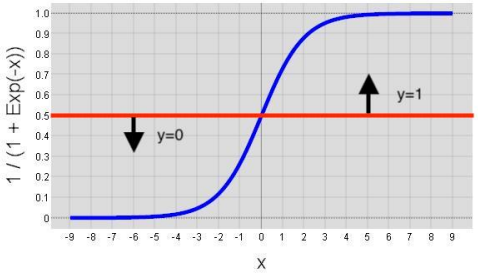
- The ROC curve and area under the curve (AUC)
- Multinomial Logistic Regression

Outline

Motivation: Threshold for Logistic Regression

For example, in logistic regression,

$$P(y = 1 \mid \boldsymbol{x}) = \sigma(\beta^T \boldsymbol{x}) \geq \text{threshold}$$



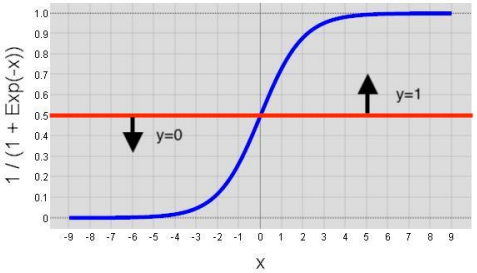
Motivation: Threshold for Logistic Regression

For example, in logistic regression,

$$P(y = 1 \mid \boldsymbol{x}) = \sigma(\beta^T \boldsymbol{x}) \geq \text{threshold}$$

$\sigma(\beta^T \boldsymbol{x}) \geq 0.5$ is one threshold to generate a binary prediction y

$$y = \begin{cases} 1 \text{ if } \sigma(\beta^T \boldsymbol{x}) \geq 0.5 \\ 0 \text{ if } \sigma(\beta^T \boldsymbol{x}) < 0.5 \end{cases}$$



but choosing the threshold can be tricky for imbalanced classes.

TPR and FPR

		predicted labels: positive (1)	predicted labels: negative (0)
true labels	positive (1)	true positive (TP)	false negative (FN)
	negative (0)	false positive (FP)	true negative (TN)

- True positive rate, $TPR = \frac{TP}{TP+FN}$
- False positive rate, $FPR = \frac{FP}{FP+TN}$

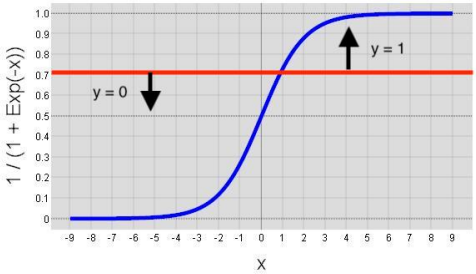
Motivation: Threshold for Logistic Regression

For example, in logistic regression,

$$P(y = 1 \mid \boldsymbol{x}) = \sigma(\beta^T \boldsymbol{x})$$

$$y = \begin{cases} 1 \text{ if } \sigma(\beta^T \boldsymbol{x}) \geq 0.7 \\ 0 \text{ if } \sigma(\beta^T \boldsymbol{x}) < 0.7 \end{cases}$$

Choosing higher threshold can lead to less false positives.



TPR and FPR

Example: Suppose you build a logistic regression classifier to predict credit card fraud from recent transactions.

Customers would rather be warned even when things are OK than let actual fraud be missed.

This means we're willing to accept a high _____ in order to secure a high _____ by choosing a _____ threshold.

A. TPR, FPR, high
B. FPR, TPR, low

TPR and FPR

Example: Suppose you build a logistic regression classifier to predict credit card fraud from recent transactions.

Customers would rather be warned even when things are OK than let actual fraud be missed.

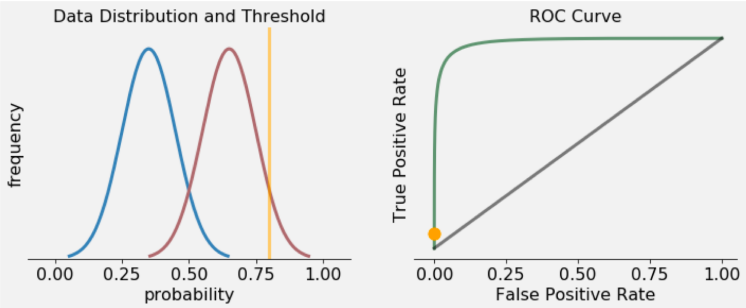
This means we're willing to accept a high _____ in order to secure a high _____ by choosing a _____ threshold.

A. TPR, FPR, high

B. FPR, TPR, low

The answer is B.
A ROC Curve gives us a visual way to evaluate suitable thresholds to fit our needs.

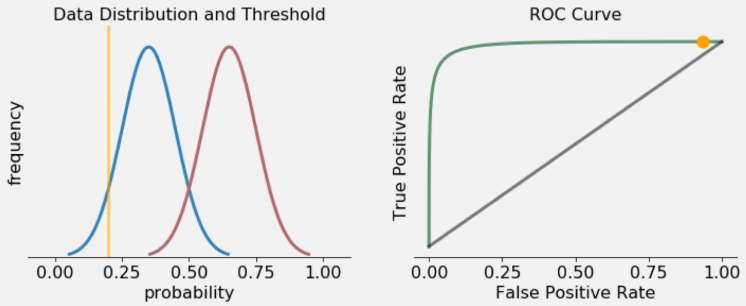
The ROC curve



A ROC Curve is a plot of FPR (horizontal) vs. TPR (vertical) for all possible threshold values.

Shows how a model would perform at all thresholds simultaneously, rather than looking at each threshold individually.

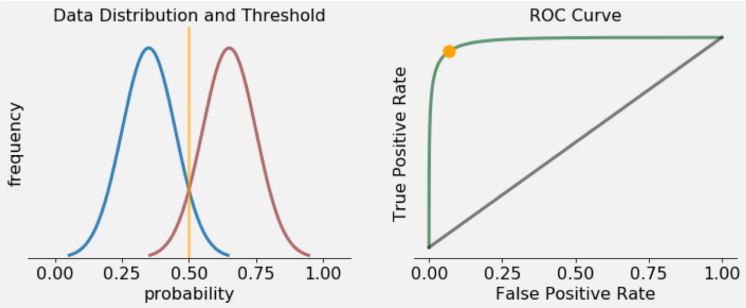
The ROC curve



A ROC Curve is a plot of FPR (horizontal) vs. TPR (vertical) for all possible threshold values.

Shows how a model would perform at all thresholds simultaneously, rather than looking at each threshold individually.

The ROC curve

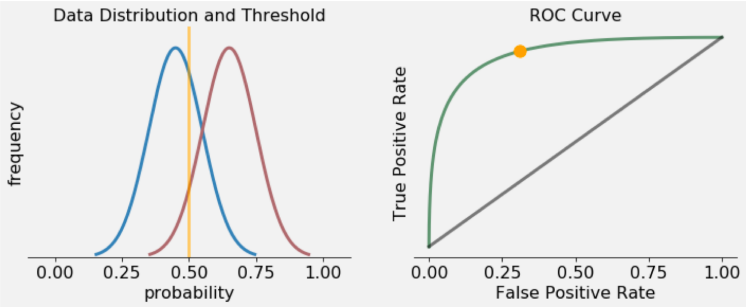


The threshold gives the parameterization of the ROC curve (i.e., it moves the dot).

When the threshold separates the two classes fairly well, the curve is far away from the diagonal.

What happens if we can't separate the classes very well?

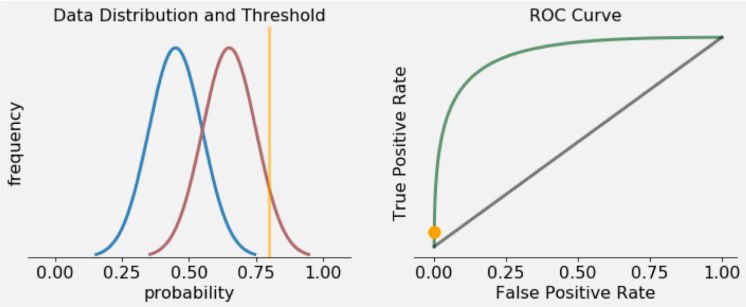
The ROC curve



Now we're not doing so well at separating the classes.

The ROC curve starts bending towards the center.

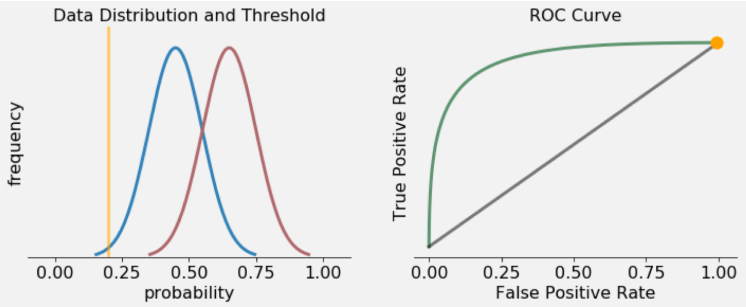
The ROC curve



Now we're not doing so well at separating the classes.

The ROC curve starts bending towards the center.

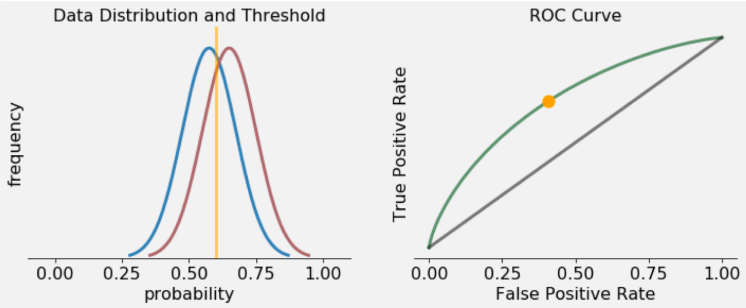
The ROC curve



Now we're not doing so well at separating the classes.

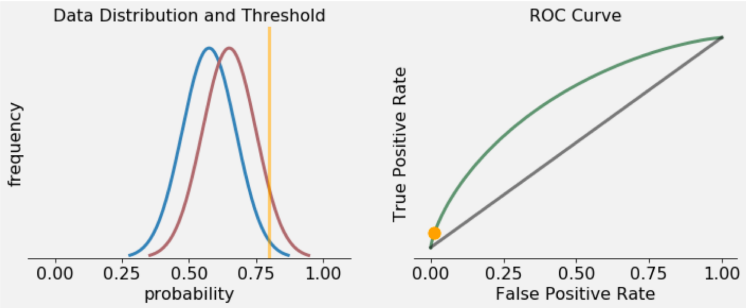
The ROC curve starts bending towards the center.

The ROC curve



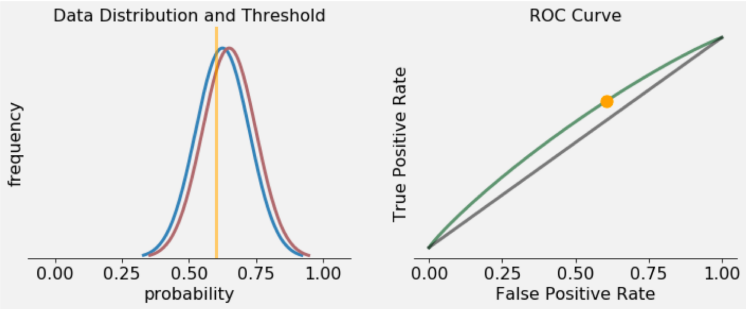
And as we do a poorer job of separating the classes, the curve continues to bend.

The ROC curve



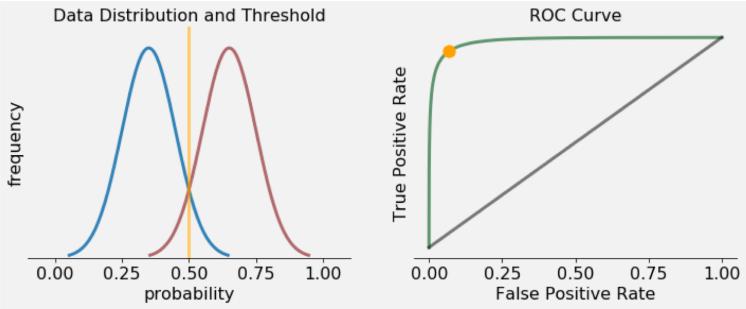
And as we do a poorer job of separating the classes, the curve continues to bend.

The ROC curve



And if we do a terrible job, the curve approaches the random chance line, indicating that our classifier is not much better than a random guess.

The ROC curve



The ROC curve addresses the cases when we're worried about FPs and TPs simultaneously.
But, if you want a single number, evaluating how the model will do in all cases You can compute the AUC (Area under the ROC curve).

ROC-AUC comparisons

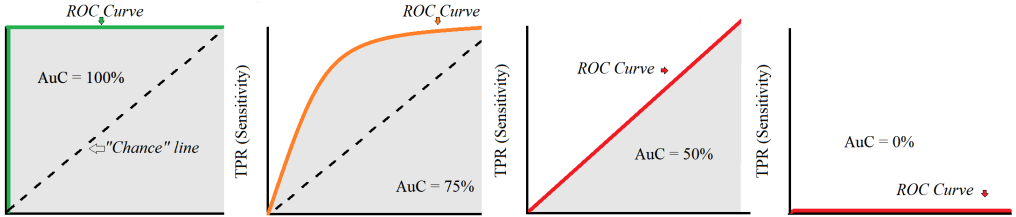


Image from `https://docs.paperspace.com/machine-learning/wiki/auc-area-under-the-roc-curve`

To compare two models, plot their ROC curves on the same axes.
If one encloses the other, then it's better on both ends of the spectrum, and has higher AUC.

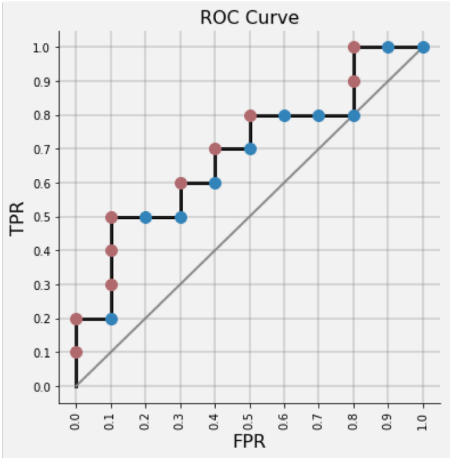
Constructing a ROC curve

You need a classifier that is able to rank examples by predicted score $\hat{p} = \sigma(\beta^T \boldsymbol{x})$.

- Order all examples by prediction confidence
- Move threshold to each point, one at a time
- If point is positive (P), move vertically (↑TP)
- If point is negative (N), move horizontally (↑FP)

#	c	$\hat{p}$	#	c	$\hat{p}$
1	P	0.90	11	P	0.40
2	P	0.80	12	N	0.39
3	N	0.70	13	P	0.38
4	P	0.60	14	N	0.37
5	P	0.55	15	N	0.36
6	P	0.54	16	N	0.35
7	N	0.53	17	P	0.34
8	N	0.52	18	P	0.33
9	P	0.51	19	N	0.30
10	N	0.50	20	N	0.10

Constructing a ROC curve



#	c	$\hat{p}$	#	c	$\hat{p}$
1	P	0.90	11	P	0.40
2	P	0.80	12	N	0.39
3	N	0.70	13	P	0.38
4	P	0.60	14	N	0.37
5	P	0.55	15	N	0.36
6	P	0.54	16	N	0.35
7	N	0.53	17	P	0.34
8	N	0.52	18	P	0.33
9	P	0.51	19	N	0.30
10	N	0.50	20	N	0.10

ROC curve

ROC cares both about TPR and FPR, so it values both positive examples and negative examples.
If only positive examples are important, one can plot precision and recall curve.

Outline

Multiclass Classification: Multinomial Logistic Regression

So logistic regression is for binary classification i.e., $y = 0, 1$.

How can we extend beyond binary classification i.e., $y = 0, 1, ..., k$ or multiclass classification?

Multiclass Classification: Multinomial Logistic Regression

Given a weight z_i assigned to class i, we use **softmax** function,

$$\text{softmax}(z_i) = \frac{\exp(z_i)}{\sum_{j=0}^{k} \exp(z_k)}$$

to calculate $P(y = i \mid \boldsymbol{x})$

Characteristics of softmax function:

- $\text{softmax}(z_i)$ is in the range from 0 to 1
- $\sum_i^k \text{softmax}(z_i) = 1$

An example of z from [3],
If $z = [z_0, ..., z_5] = [0.6, 1.1, -1.5, 1.2, 3.2, -1.1]$,
then $\text{softmax}(z) = [0.055, 0.090, 0.006, 0.099, 0.74, 0.010]$

Multiclass Classification: Multinomial Logistic Regression

For mutinomial logistic regression: $z_i = \beta_i^T \boldsymbol{x}$

For each class i, we keep track of separate parameter β_i,

$$P(y = i \mid \boldsymbol{x}) = \frac{\exp(\beta_i^T \boldsymbol{x})}{\sum_{j=0}^{k} \exp(\beta_j^T \boldsymbol{x})}$$

Is this related to logistic regression or logistic function?

Logistic Function as a Special Case of Softmax Function [1, 2]

When the number of classes $k = 2$ and if $-\beta^T = (\beta_0^T - \beta_1^T)$, then

$$P(y = 1 \mid \boldsymbol{x}) = \frac{\exp(\beta_1^T \boldsymbol{x})}{\exp(\beta_0^T \boldsymbol{x}) + \exp(\beta_1^T \boldsymbol{x})} \tag{1}$$

$$= \frac{1}{\exp(\beta_0^T \boldsymbol{x}) \exp(-\beta_1^T \boldsymbol{x}) + 1} \tag{2}$$

$$= \frac{1}{\exp((\beta_0^T - \beta_1^T)\boldsymbol{x}) + 1} \tag{3}$$

$$= \frac{1}{\exp(-\beta^T \boldsymbol{x}) + 1} = \sigma(\beta^T \boldsymbol{x}) \tag{4}$$

softmax function is a generalization of logistic function

Scikit-learn Implementation of Logistic Regression

But there seems more option in the “multiclass” argument:

multi_class : *{‘auto’, ‘ovr’, ‘multinomial’}, default=‘auto’*

If the option chosen is ‘ovr’, then a binary problem is fit for each label. For ‘multinomial’ the loss minimised is the multinomial loss fit across the entire probability distribution, *even when the data is binary*. ‘multinomial’ is unavailable when solver=’liblinear’. ‘auto’ selects ‘ovr’ if the data is binary, or if solver=’liblinear’, and otherwise selects ‘multinomial’.

`https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

“ovr” is the abbreviation for “one versus rest” (or “one against all”) which we will talk about it in the next lecture...

References

[1] Softmax vs Sigmoid function in Logistic classifier? https://stats.stackexchange.com/questions/233658/softmax-vs-sigmoid-function-in-logistic-classifier/254071. Accessed on 02.14.2021.

[2] UFLDL Tutorial. http://deeplearning.stanford.edu/tutorial/supervised/SoftmaxRegression. Accessed on 02.14.2021.

[3] Dan Jurafsky and James H. Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 3rd edition. https://web.stanford.edu/~jurafsky/slp3/5.pdf, 2020.