Performance Evaluation

UW ECE 657A - Core Topic

Mark Crowley

Lecture Outline

1. Classification Confidence
   - Confusion Matricies
2. Measuring Parameter/Threshold Tradeoffs
   - Receiver Operating Characteristic (ROC) Curve
   - Precision-Recall Curve
3. Other Performance Measures
4. Underfitting, Overfitting and Capacity

Classification Confidence

For a binary classification problem (ie. $C = \{C_0, C_1\}$) our decision rule could be

$$\delta(x) = \mathcal{I}(f(x) > \tau)$$

where:

- $f(x)$ is a measure of confidence that $x \in C_1$. This could be a probability, a distance or other function
- τ (*"tau"*) is a **threshold parameter** that is used to make a decision on which class to assign to each x.

For different values of τ we will get a different number of $x_i \in C_1$. **todo:** copy confusion matrix from hypoth testing discussion, retrieval vs calssification terminology YAH

Using a Confusion Matrix

For a particular value of τ we can build a **confusion matrix**:

		True Value	
		1	0
Estimated	1	**TP**	**FP**
	0	**FN**	**TN**
	Sum	$N_+ = TP + FN$	$N_- = FP + TN$

TP: True Positive

FP: False Positive (False alarm)

TN: True Negative

FN: False Negative (Missed detection

Using a Confusion Matrix

For a particular value of τ we can build a **confusion matrix**:

		True Value	
		1	0
Estimated	1	**TP**	**FP**
	0	**FN**	**TN**
	Sum	$N_+ = TP + FN$	$N_- = FP + TN$

- $N = TP + FP + FN + TN$
- $\hat{N}_+ = TP + FP$
- $\hat{N}_- = FN + TN$

Using a Confusion Matrix

For a particular value of τ we can build a **confusion matrix**:

		True Value	
		1	0
Estimated	1	**TP**	**FP**
	0	**FN**	**TN**
	Sum	$N_+ = TP + FN$	$N_- = FP + TN$

Notes:

- For more than one class you could build this for each class, whether the point is in the class or not.

The False Positive vs False Negative Tradeoff

From this table we can also compute various success and error rates:

		True Value	
		1	0
Estimated	1	**TPR**$= TP/N_+$	**FPR**$= FP/N_-$
	0	**FNR**$= FN/N_+$	**TNR**$= TN/N_-$

TPR: True Positive Rate (Sensitivity, Recall, Hit rate)

FPR: False Positive Rate (False alarm rate, Type I Error)

TNR: True Negative Rate (Miss Rate, Type II Error)

FNR: True Negative Rate (Specificity)

Remember, this depends on τ, so how do we find the *best* value of τ?

Receiver Operating Characteristic (ROC) Curve

- ROC Originally conceived during WW II to assess the performance of radar systems
- If we apply our decision rule $\delta(x)$ for a range of τ then we can draw a curve of any of the success/error rates.
- If we plot TPR vs FPR we get the ROC Curve
- Say we have two classifiers or distributions. One for C_0 and another for C_1

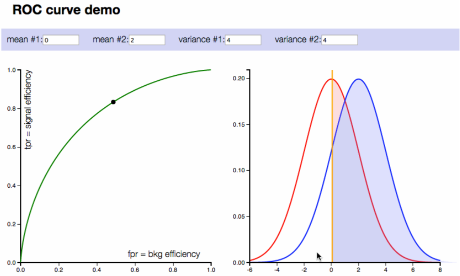
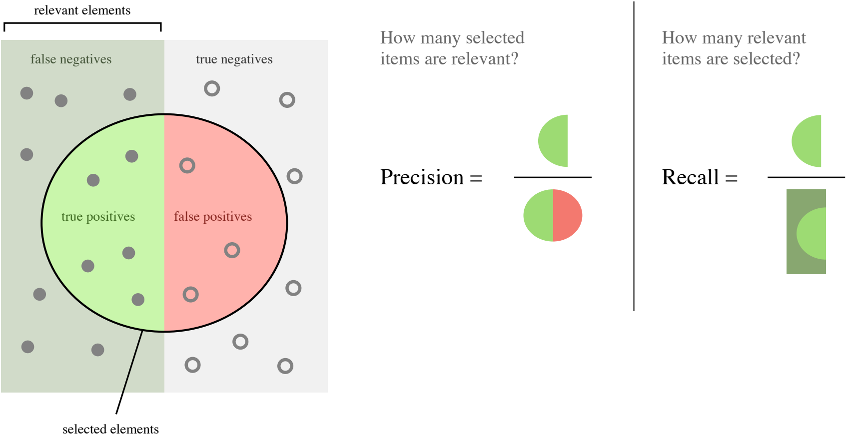


Figure: Demo of a ROC curve at `http://arogozhnikov.github.io/2015/10/05/roc-curve.html`

Precision vs. Recall



[Image from `https://en.wikipedia.org/wiki/Precision_and_recall`]

Another Confusion Matrix

We can also create a confusion matrix using our estimated count of positives and negatives:

		True Value	
		1	0
Estimated	0	$TP/\hat{N}_+$=PPV=**precision**	$FP/\hat{N}_+$=FDP
	1	$= FN/\hat{N}_-$	**TNR**$= TN/\hat{N}_-$

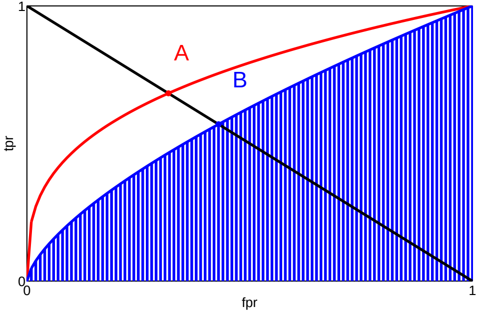
Precision measures what fraction of our detections are actually positive.
Recall measures what fraction of all the positives that we actually detected.

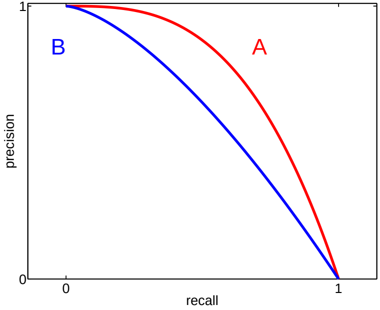
Precision-Recall Curve

- If we plot precision P vs recall R as we vary τ then we get a **precision-recall curve** which can often show us different performance behaviour from the ROC curve.
- The P-R only uses statistics based on TP, so the curve is useful when there is a very small number of positive cases in your classifier or when the number of negatives could scale based on some parameter.
- Curve bends the other way. **todo:** more analysis of P-R curve?

ROC vs PR Curves

Line A is better than line B in both curves.





Other Error Measures

- Accuracy : $\frac{TP+TN}{P+N}$
- Error: $\frac{FN+FP}{P+N}$
- Precision: $\frac{TP}{TP+FP}$
- Recall/Sensitivity: $\frac{TP}{TP+FN}$
- F-measure (F1-score) : $\frac{2*\texttt{Precision}*\texttt{Recall}}{\texttt{Precision}+\texttt{Recall}}$

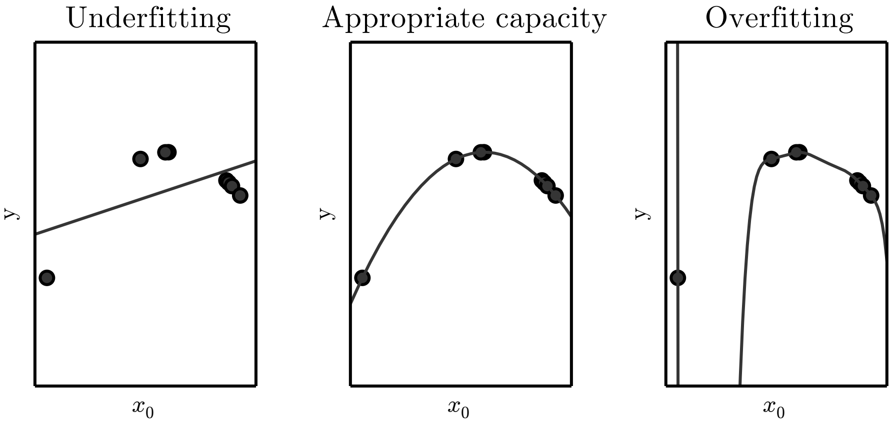
Underfitting, Overfitting and Capacity

- **Generalization**: is the ability to perform well on previously unobserved inputs.
- **Generalization Error/Test Error**: the expected error on new inputs.

Another way to see our goal then:

- Avoid Underfitting: Make the training error small
- Avoid Overfitting: Make the gap between training and test error small

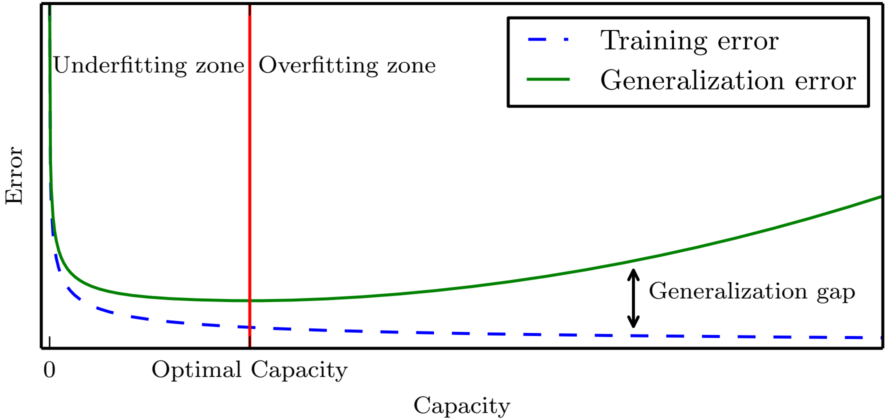
Underfitting vs. Overfitting



(Goodfellow 2016)

Capacity

- **Capacity**: ability of a model to fit a wide variety of functions



(Goodfellow 2016)

Quantifying Capacity

- Quantifying capacity precisely is hard
- **VC dimension**: measures the capacity of a binary classifier.
    - the largest possible value m for which there exists a training set of m different points in X that the classifier can label arbitrarily.
    - this is very hard to define or use in practice, but makes for some good proofs for performance bounds on classification algorithms.
- To get near the highest end of capacity we need to go to **non-parametric** models.