Session 5a:
Model Validation Techniques

Marek J. Drużdżel

Wydział Informatyki

Politechnika Białostocka

m.druzdzel@pb.edu.pl

http://aragorn.wi.pb.bialystok.pl/~druzdzel/

Course schedule

Day 1

Session 1: Introduction to probabilistic (Bayesian) modeling and inference

Session 2: Bayesian networks

Session 3: Building Bayesian networks

Session 4: Hands-on exercises (Bayesian networks)

Day 2

Session 5: **Learning: structure/causal discovery, parameter learning, model validation techniques**

Session 6: **Hands-on exercises (learning)**

Session 7: **Decision analysis: expected utility theory, utility elicitation, influence diagrams**

Session 8: **Hands-on exercises (decision analysis)**

Session overview

- Introduction
- Cross-validation
- Basic model quality measures: accuracy, sensitivity, specificity, confusion matrices
- ROC curves and AUC
- Calibration curves
- Software demo
- Concluding remarks

What I want you to know after this session

- Appreciate the need for model validation
- Understand the representative sample-based approaches to validation by means of data
- Appreciate the need for cross-validation
- Know the concepts of accuracy, sensitivity, specificity, ROC, AUC, and calibration
- Be able to perform validation by means of software

The task

"I can prove it or disprove it! What do you want me to do?"

Introduction: The need for verification

The fundamental question:

How do we know that the knowledge extracted from data is worth anything?

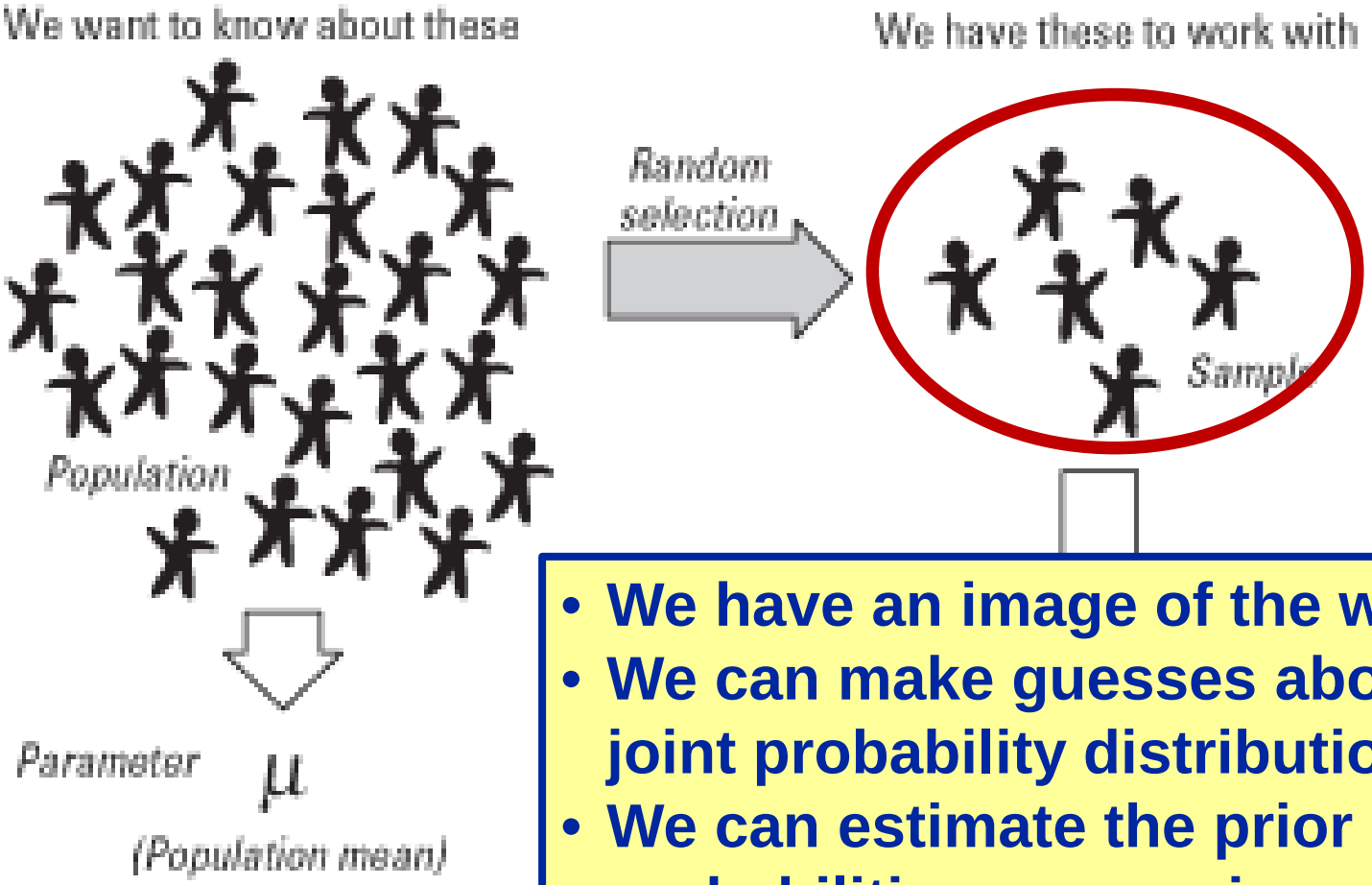
http://www.ehow.com/how_7897502_evaluate-higher-order-questions-answers.html

Introduction: The need for verification

Many possible (often problem-dependent) answers, e.g.,

- ***"I'm just reporting what I see in the data"*** **(How do you know that what you see is what is there ☺?)**
- ***"My model performs well in practice"*** **(What does it mean to "perform well"?)**
- ***"I can provide a measure of reliability of the extracted knowledge"*** **(usually in terms of statistical parameters, such as *p*-value or confidence interval)**
- ***"I have confirmed the discovery independently"*** **(e.g., a causal hypothesis by manipulating the world and observing correctness of the model's predictions)**

"Representative sample"-based approaches



- **We have an image of the world!**
- **We can make guesses about the joint probability distribution.**
- **We can estimate the prior probabilities over various hypotheses.**

http://www.cliffsnotes.com/study

Cross-Validation

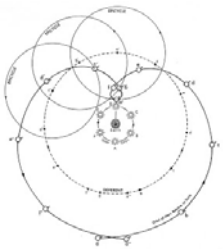
Cross-validation: The idea

Testing a model on the same data that we used for training does not seem fair.

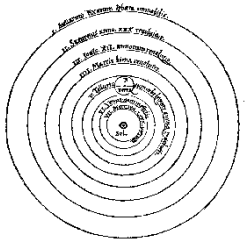
It would be like training students using precisely the questions that we are going to ask them at an exam.
What is the best strategy?
Memorize the answers!
How will the students perform on questions that they have never seen?
Quite possibly poorly.

Cross-validation: The idea

- Testing a model on the same data that we used for training it does not seem fair.
- It will favor most complex models that fit the data best (e.g., the best strategy will be to learn all training cases by hard, no matter what it takes).
- Simpler models may actually fit future instances of data better than complex models.



Ptolemy's model



Copernicus' model

Cross-validation prevents over-fitting

Cross-validation: Hold-out method

Cross-validation is a technique for assessing how the results will generalize to an independent data set.

Used in settings where the goal is prediction and estimates the practical accuracy of the model.

- Divide the data into two disjoint sets: (1) training set and (2) test set (a.k.a. validation set). The size of the two subsets is a matter of decision and usually depends on the size of the data set.
- Learn from the training set and validate the results on the test set.

Simple and effective ☺.

What are the disadvantages of this approach?

It wastes data that could have been used for learning ☹.

With small data sets, possibly subject to luck/coincidence (when the test set has high variance) ☹.

Cross-validation: k-fold Cross-validation

"k-fold" cross-validation

Divides the data set into k (roughly) equal-size parts, uses k-1 of these for training and the remaining one for testing, repeats this k times, averages the results over the folds.

Iteration 1	Iteration 2	Iteration 3	...	Iteration k
Fold 1	Fold 1	Fold 1	...	Fold 1
Fold 2	Fold 2	Fold 2	...	Fold 2
Fold 3	Fold 3	Fold 3	...	Fold 3
...	...	...	...	...
Fold k	Fold k	Fold k	...	Fold k

Training

Testing

This reduces variability in the result at the expense of more computation.

"Leave-One-Out" cross-validation

Uses effectively n-1 instances for training and tests the model on all n instances, one at a time (an extreme case of k-fold, k=n).

Cross-validation: Bootstrap cross-validation

Bootstrap cross-validation involves multiple repetitions of the hold-out method. Each repetition involves selecting (with replacement) a new training set from among all records. Those records that do not appear in the training set become the test set.

Because it can (and should!) be repeated many times, it may involve a lot more computation than k-fold cross validation ☹.

It can outperform cross-validation in some cases ☺.

Accuracy, Sensitivity, Specificity, Confusion Matrix

Accuracy

Count how many instances identified (classified, recognized, guessed) correctly

Problems with accuracy alone:

- **Sensitivity to the base rate.**
- **Let prevalence of cancer be 10 in a 1000**
- **A model that always guesses “no cancer” will have 99% accuracy but will miss all of the cancers.**

Need to look at more details!

http://www.opsrules.com/supply-chain-optimization-blog/bid/282916/Why-You-Shouldn-t-Waste-Effort-Improving-Forecast-Accuracy

Sensitivity and Specificity

Sensitivity and specificity

Sensitivity and specificity are statistical measures of the performance of a binary classification test, also known in statistics as classification function.

Sensitivity (also called recall rate in some fields) measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition).

Specificity measures the proportion of negatives which are correctly identified (e.g. the percentage of healthy people who are correctly identified as not having the condition).

These two measures are closely related to the concepts of type I and type II errors. A perfect predictor would be described as 100% sensitivity (i.e., predict all people from the sick group as sick) and 100% specificity (i.e., not predict anyone from the healthy group as sick).

In practice, however, there are no perfect predictors.

http://en.wikipedia.org/wiki/Specificity_(statistics)

Sensitivity and specificity: Definitions

Sensitivity

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

$$= \text{probability of a positive test given that the patient is ill}$$

Specificity

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

$$= \text{probability of a negative test given that the patient is well}$$

http://en.wikipedia.org/wiki/Specificity_(statistics)

Sensitivity and specificity: Relationship among terms

The fecal occult blood (FOB) screen test used in 2,030 people to look for bowel cancer

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	True Positive	False Positive (Type I error)	Positive predictive value = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	False Negative (Type II error)	True Negative	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	Specificity = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

http://en.wikipedia.org/wiki/Specificity_(statistics)

Accuracy=(TN+TP)/(TN+TP+FN+FP)=1840/2030=90.6%

Sensitivity and specificity: Example

The fecal occult blood (FOB) screen test used in 2030 people to look for bowel cancer

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	True Positive (TP) = 20	False Positive (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test Outcome Negative	False Negative (FN) = 10	True Negative (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	**Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

http://en.wikipedia.org/wiki/Specificity_(statistics)

Sensitivity and specificity: Example

Related calculations

False positive rate (α) = type I error = 1 − specificity = FP / (FP + TN) = 180 / (180 + 1820) = 9%

False negative rate (β) = type II error = 1 − sensitivity = FN / (TP + FN) = 10 / (20 + 10) = 33%

Power = sensitivity = 1 − β

Likelihood ratio positive = sensitivity / (1 − specificity) = 66.67% / (1 − 91%) = 7.4

Likelihood ratio negative = (1 − sensitivity) / specificity = (1 − 66.67%) / 91% = 0.37

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	True Positive (TP) = 20	False Positive (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test Outcome Negative	False Negative (FN) = 10	True Negative (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	**Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

Hence, with large numbers of false positives and few false negatives, a positive FOB screen test is in itself poor at confirming cancer (PPV = 10%) and further investigations must be undertaken; it did, however, correctly identify 66.7% of all cancers (the sensitivity). However as a screening test, a negative result is very good at reassuring that a patient does not have cancer (NPV = 99.5%) and at this initial screen correctly identifies 91% of those who do not have cancer (the specificity).

http://en.wikipedia.org/wiki/Specificity_(statistics)

Confusion Matrix

Confusion matrix

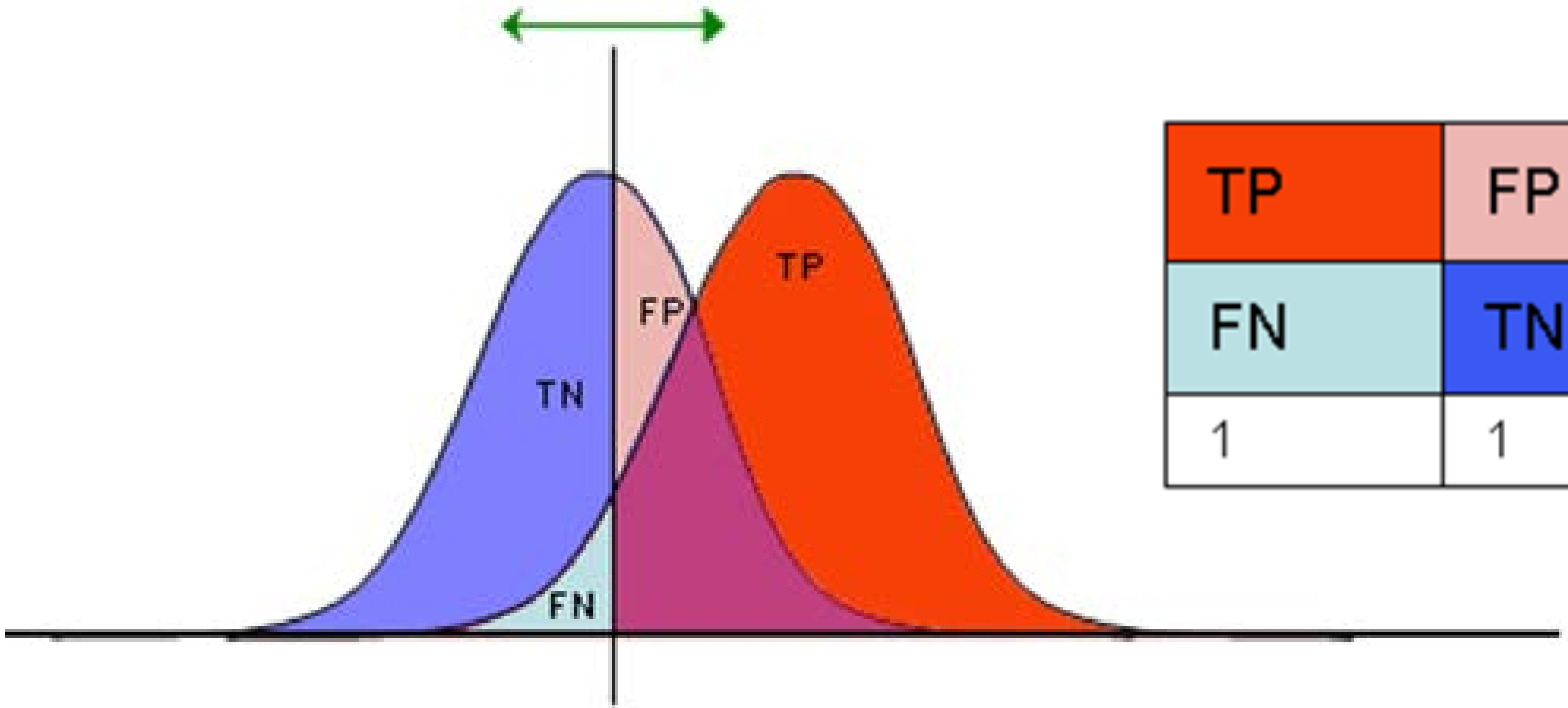
The same thing as what we saw before but used with reference to the model's predictions and the true state of the World.

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	True Positive (TP) = 20	False Positive (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test Outcome Negative	False Negative (FN) = 10	True Negative (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	**Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

http://en.wikipedia.org/wiki/Specificity_(statistics)

ROC (Receiver Operating Characteristic) Curves

ROC curve (the common sense)

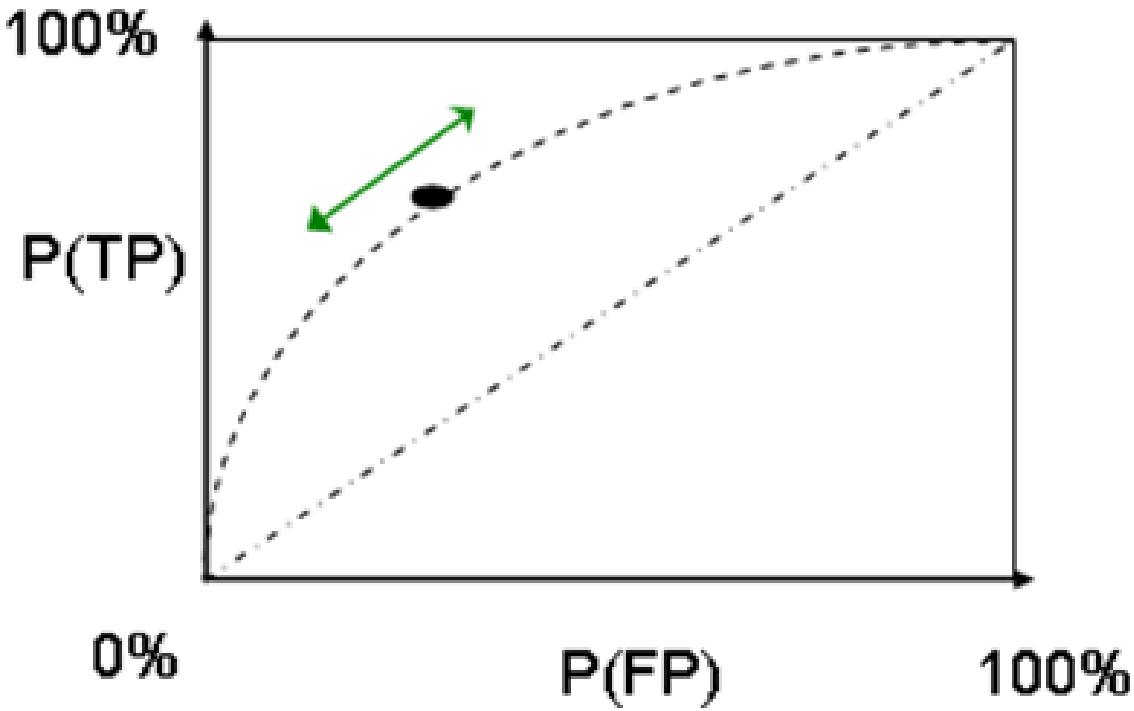


TP	FP
FN	TN
1	1

- Please note that very often we need to make a compromise between sensitivity and specificity: Higher sensitivity means lower specificity and vice versa.
- Setting the threshold is a matter of decision.
- The threshold that we decide to adopt will determine the parameters of our test (i.e., true/false positive and true/false negative rates).

http://en.wikipedia.org/wiki/Receiver_operating_characteristic

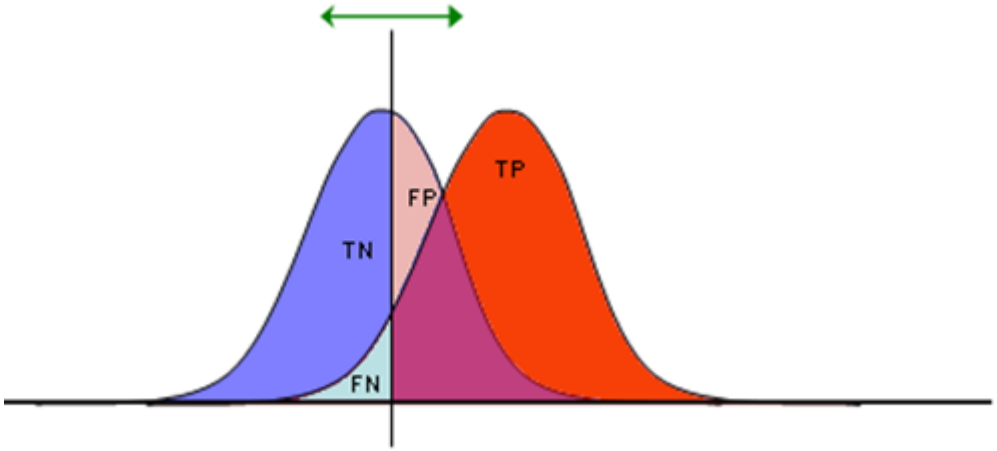
ROC curve (the common sense)



- As we move the threshold, we change the values of sensitivity and specificity. The plot of all possible values of these two parameters gives us an interesting characterization of the test (classification system, receiver, etc.)

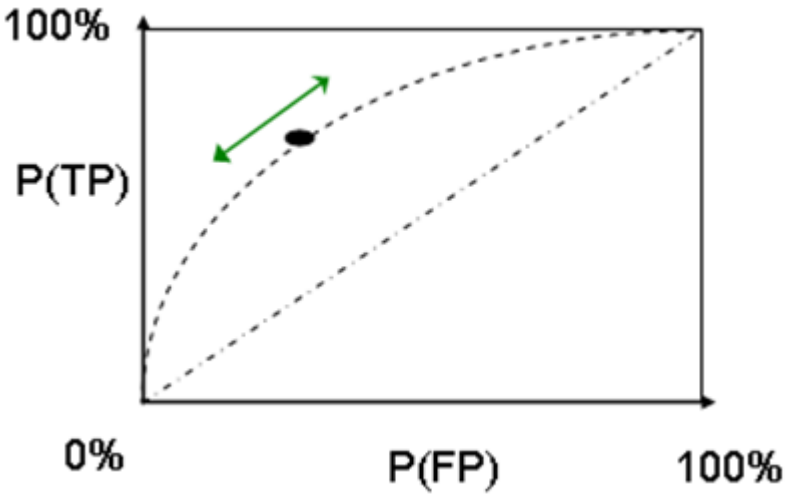
http://en.wikipedia.org/wiki/Receiver_operating_characteristic

ROC curve (the common sense)



TP	FP
FN	TN
1	1

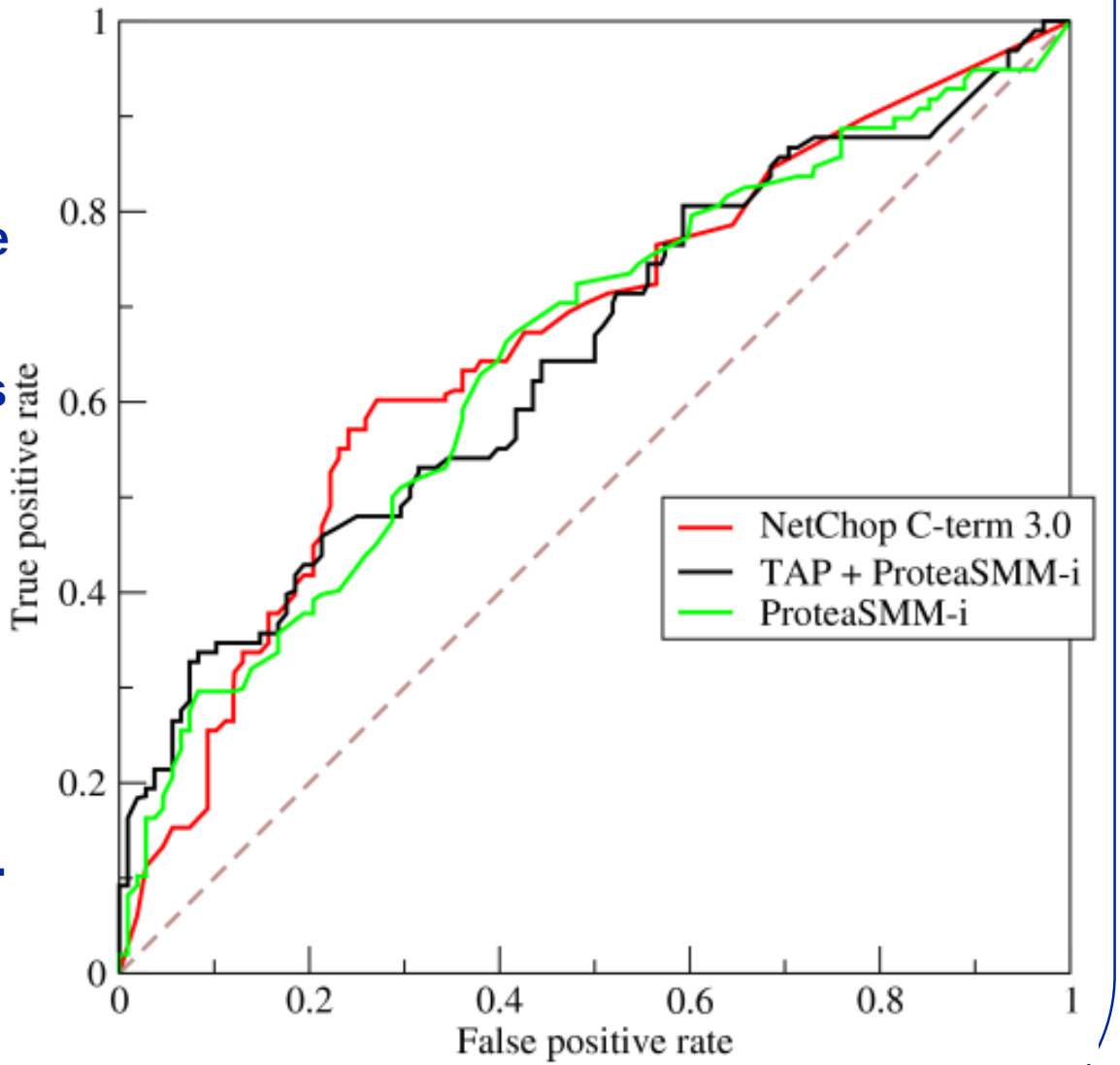
- Plots like the one on the right-hand side are called ROC (Receiver Operating Characteristics) Curves
- They are a way of characterizing the quality of the detection system



http://en.wikipedia.org/wiki/Receiver_operating_characteristic

ROC curve

- Originates from the signal detection theory
- A graphical plot that illustrates the performance of a **binary** classifier system as its discrimination threshold is varied.
- Created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate.

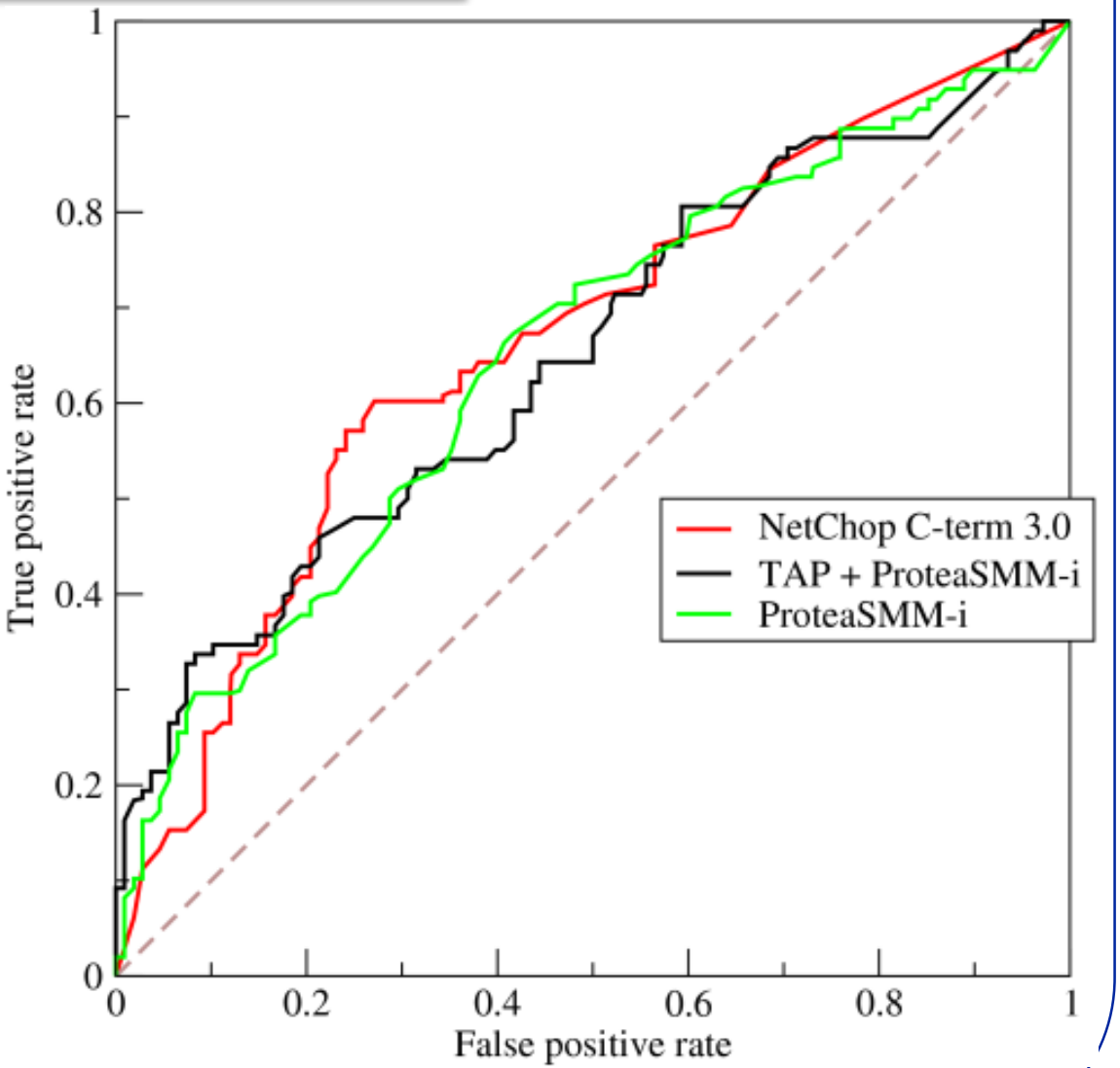


http://en.wikipedia.org/wiki/Receiver_operating_characteristic

AUC: Area Under the (ROC) Curve

AUC: Area Under the (ROC) Curve

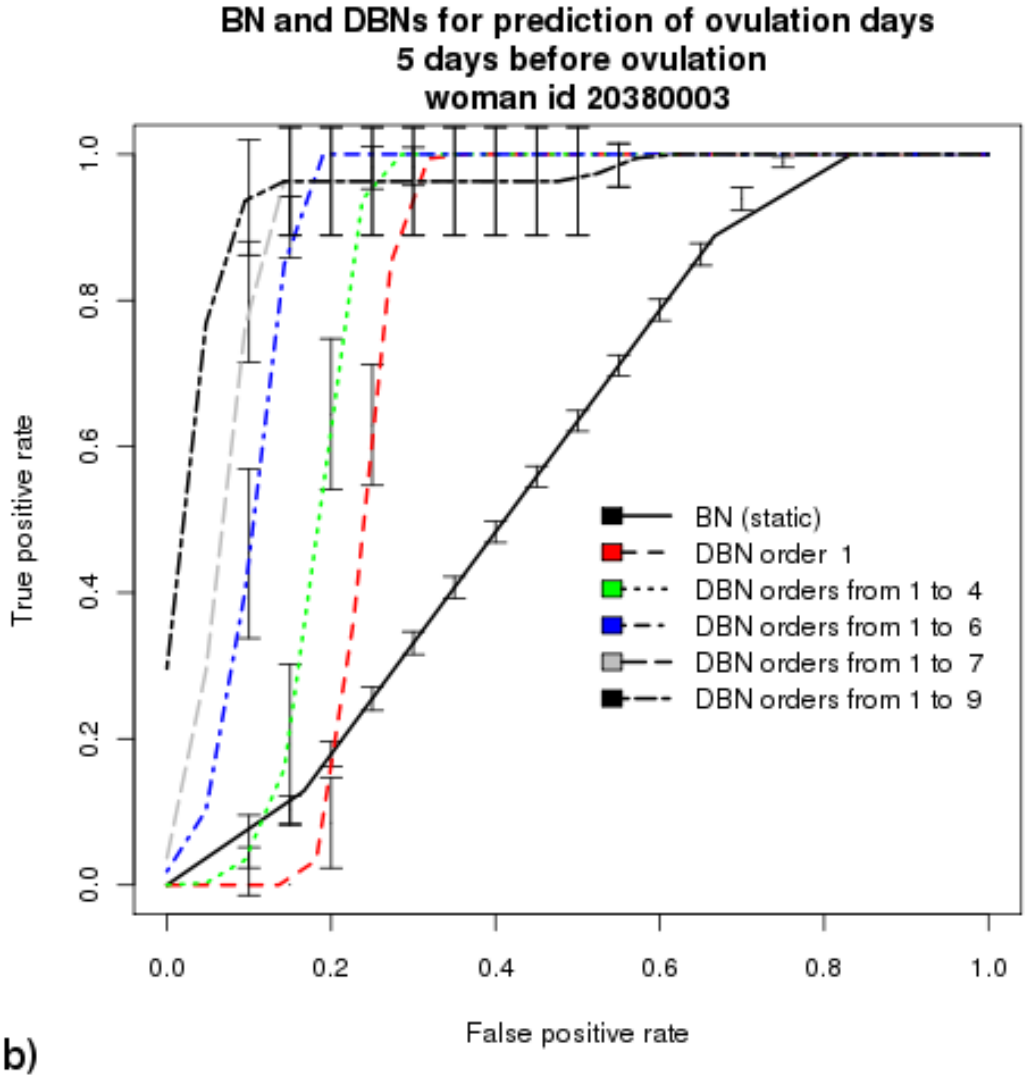
- AUC is a way of characterizing a “receiver” by means of one number
- It captures the intuitive idea that an ROC that is higher is better.
- A perfect ROC curve will go through the point (0,1). The area under it will be 1.0.



http://en.wikipedia.org/wiki/Receiver_operating_characteristic

AUC: Area Under the (ROC) Curve

AUC does not always indicate the best model



b)

Calibration

Elements of decision theory

The theoretically sound way of making decisions under uncertainty

- We need to consider uncertainty and preferences. These are measured in terms of probability and utility respectively.
- Probability is a measure of uncertainty.
- Utility is a measure of preference that combines with probability as mathematical expectation.

Example decision

http://www.fox7austin.com/weather/69360832-story

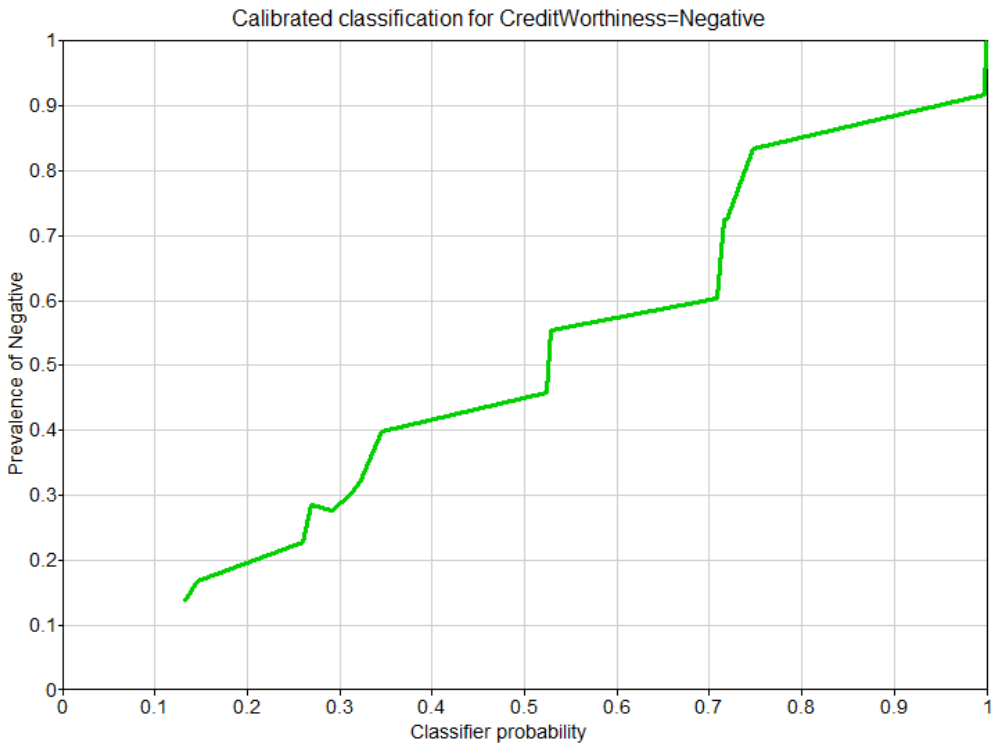
- Should we carry an umbrella?
- When is the forecast good?

Calibration

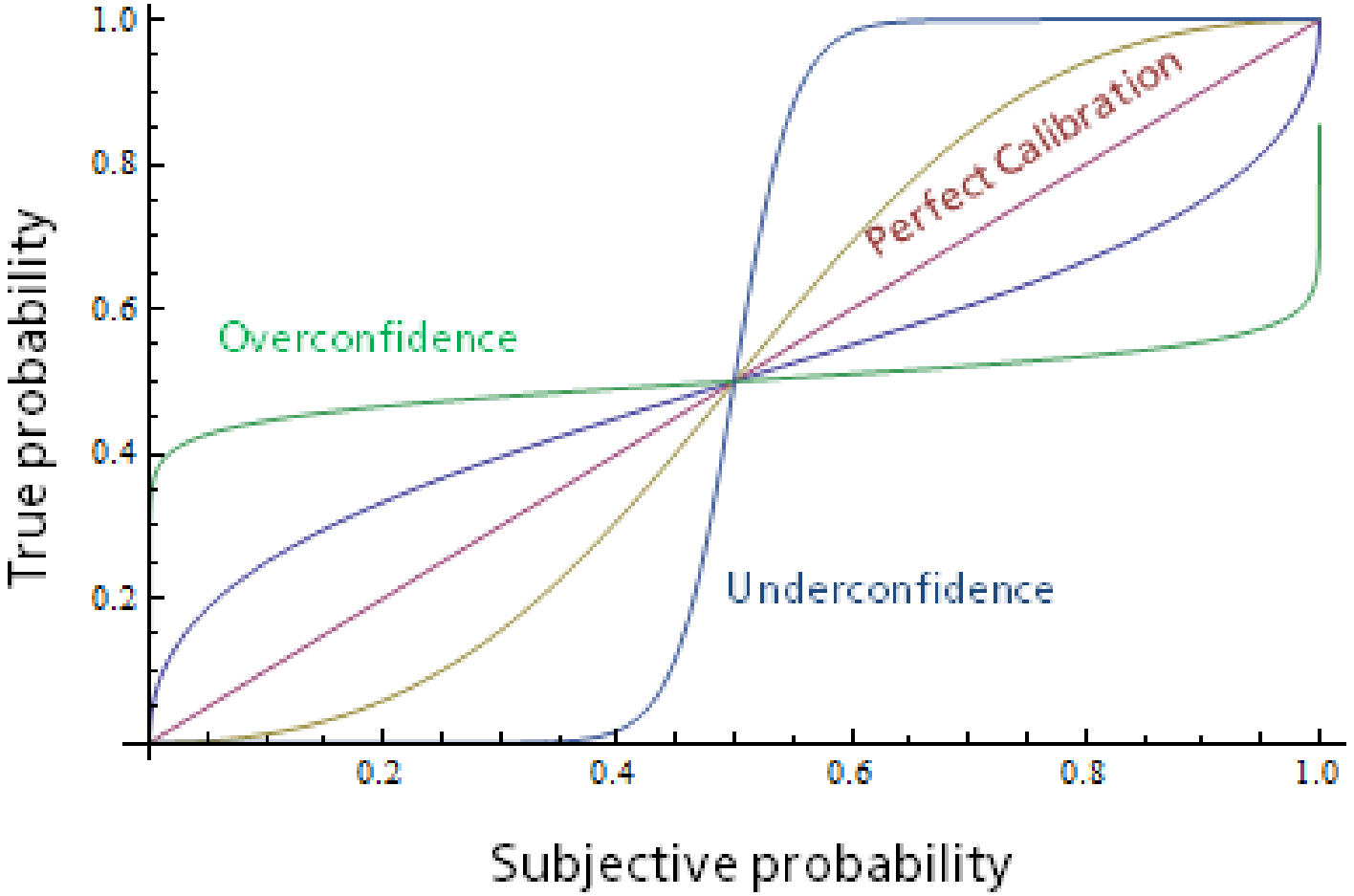
The question here is: Is my model producing accurate probabilities?

We plot the frequencies observed in the data (y axis) against the probabilities calculated by the system (x axis).

Various tricks to smoothen out the curve.

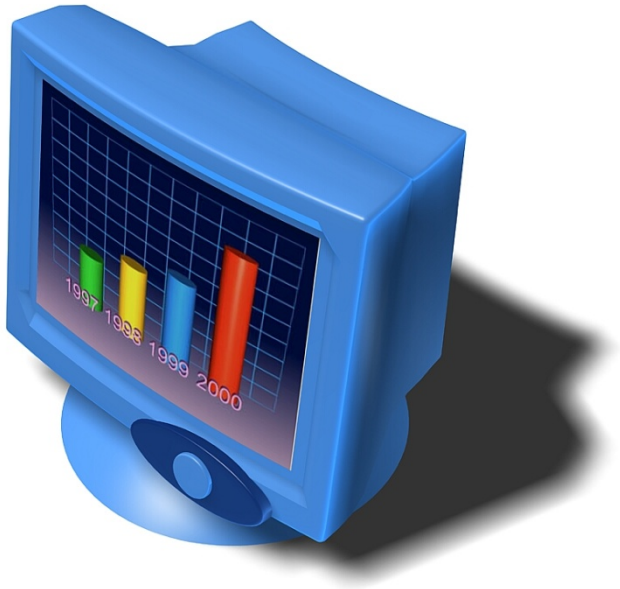


Calibration: Overconfidence and Underconfidence



http://3.bp.blogspot.com/-lmpGS0cqvuw/VDxhem5qTFI/AAAAAAAAACU/qu0hVUn9PBQ/s1600/20141014-Calibration.png

The reminder of this session



Concluding remarks

- Reality is a great check on every activity ☺.
- Verification is critical for every model and theory, including models and theories derived from data.
- Statistics is again a guiding light in this respect.
- There are a variety of approaches to verification and testing, cross-validation being the prominent one for data-based analysis.
- When a model produces probability, calibration is often forgotten/overlooked.