# Analiza danych: Techniki walidacji modeli

**Marek J. Drużdżel**

**Politechnika Białostocka**

**Wydział Informatyki**

*m.druzdzel@pb.edu.pl*
*http://www.wi.pb.edu.pl/~druzdzel/*

# Plan wykładu

- Wstęp
- Walidacja krzyżowa
- Podstawowe miary jakości modelu: dokładność, czułość, swoistość, macierz przekłamań
- Krzywe ROC i AUC
- Krzywe kalibracyjne
- Demo
- Uwagi końcowe

# Wprowadzenie: Potrzeba walidacji

**Podstawowe pytanie:**

## Skąd wiemy, że wiedza wydobyta z danych jest cokolwiek warta?

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

# Wprowadzenie: Potrzeba walidacji

**Różne (często zależne od problemu) odpowiedzi, np.**

- ***„Przekazuję tylko to, co widzę w danych”*** **(Skąd wiesz, że to, co widzisz, jest tym, co tam jest ☺?)**
- ***„Mój model radzi sobie dobrze w praktyce”*** **(Co to znaczy „dobrze sobie radzi”?)**
- ***„Mogę zapewnić miarę wiarygodności wydobytej wiedzy”*** **(Zwykle w postaci parametrów statystycznych, takich jak wartość *p* lub przedział ufności)**
- ***„Samodzielnie potwierdziłem odkrycie”*** **(np. poprzez potwierdzenie hipotezy przyczynowej eksperymentalnie)**

# Walidacja krzyżowa

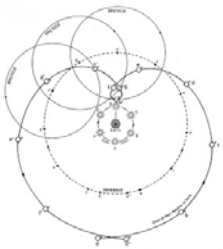
# Walidacja krzyżowa: O co chodzi

**Testowanie modelu na tych samych danych, których użyliśmy do uczenia, nie wydaje się dobrym pomysłem.**

To tak, jakby nauczać używając dokładnie tych pytań, które zadamy im na egzaminie.
Jaka jest najlepsza strategia?
Zapamiętaj odpowiedzi!
Jak uczniowie poradzą sobie z pytaniami, których nigdy nie widzieli?
Całkiem możliwe, że słabo.

# Walidacja krzyżowa: O co chodzi

- Testowanie modelu na tych samych danych, których użyliśmy do uczenia faworyzuje najbardziej złożone modele, które najlepiej pasują do danych (np. najlepszą strategią będzie dokładne nauczenie się wszystkich przypadków w zbiorze treningowym).
- Prostsze modele mogą lepiej pasować do przyszłych instancji danych niż modele złożone.



Ptolemy's model

Copernicus' model

**Walidacja krzyżowa zapobiega nadmiernemu dopasowaniu**

# Walidacja krzyżowa: Metoda „holdout”

**Walidacja krzyżowa** to technika oceny, w jaki sposób wyniki mogą zostać uogólnione na niezależny zbiór danych.

Stosowana w sytuacjach, w których celem jest przewidywanie i szacowanie praktycznej dokładności modelu.

- Dane dzieli się na dwa rozłączne zbiory: (1) zbiór uczący i (2) zbiór testowy (tzw. zbiór walidacyjny). Rozmiar tych dwóch podzbiorów jest kwestią decyzji i zwykle zależy od rozmiaru zbioru danych.
- Modelu uczy się na zbiorze treningowym, a sprawdza wyniki na zbiorze testowym.

Prosta i skuteczna metoda ☺.

Jakie są wady tego podejścia?

- Marnuje dane, które można byłoby wykorzystać do nauki ☹.
- W przypadku małych zbiorów danych, wynik ewaluacji zależy od szczęścia/zbiegu okoliczności (gdy zbiór testowy ma dużą wariancję) ☺.

# Walidacja krzyżowa: Metoda k-fold

## Metoda “k-fold”

Dzieli zbiór danych na k równych (w przybliżeniu) części, wykorzystuje k-1 z nich do uczenia, a pozostałą do testowania, powtarza to k razy, uśrednia wyniki z każdej iteracji.

| 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

Zmniejsza to zmienność wyniku kosztem większej liczby obliczeń.

## Metoda “Leave-One-Out”

Skutecznie wykorzystuje n-1 instancji do uczenia i testuje model na wszystkich n instancjach, pojedynczo (skrajny przypadek metody „k-fold”, gdy k=n).

# Dokładność, czułość, swoistość, macierz przekłamań

# Dokładność

**Liczy ile instancji zostało poprawnie zidentyfikowanych (sklasyfikowanych, rozpoznanych, odgadniętych).**

**Problemy z dokładnością:**

- **Wrażliwość na asymetrie w rozkładach zmiennych klas.**
- **Niech częstość występowania raka wynosi 10 na 1000**
- **Model, który zawsze zgaduje „nie ma raka”, będzie miał 99% dokładności, ale przeoczy wszystkie nowotwory.**

**Trzeba przyjrzeć się szczegółom!**

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

# Czułość i swoistość

# Czułość i swoistość

Czułość i swoistość to statystyczne miary jakości binarnego testu klasyfikacyjnego, znanego również w statystyce jako funkcja klasyfikacyjna.

**Czułość** mierzy odsetek faktycznie pozytywnych wyników, które zostały prawidłowo zidentyfikowane (np. odsetek chorych, u których prawidłowo zidentyfikowano daną chorobę).

**Swoistość** mierzy odsetek prawidłowo zidentyfikowanych wyników negatywnych (np. odsetek zdrowych osób, które zostały prawidłowo zidentyfikowane jako osoby nie cierpiące na daną chorobę).

Te dwie miary są ściśle powiązane z koncepcjami błędów typu I i II. Doskonały predyktor nie popełnia błędów, czyli charakteryzuje się 100% czułością (tj. określa wszystkie osoby z grupy chorych jako chore) i 100% swoistością (tj. nie charakteryzuje nikogo z grupy zdrowych jako chorego).

W praktyce jednak nie ma doskonałych predyktorów.

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

# Czułość i swoistość: Definicje

## Czułość

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

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

## Swoistość

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

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

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

# Czułość i swoistość: Związki pomiędzy pojęciami

**Badanie przesiewowe na krew utajoną w kale (FOB) zastosowane u 2.030 osób w celu wykrycia raka jelita grubego**

| | | 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 = Σ True Positive / Σ Test Outcome Positive |
| | Test Outcome Negative | False Negative (Type II error) | True Negative | Negative predictive value = Σ True Negative / Σ Test Outcome Negative |
| | | Sensitivity = Σ True Positive / Σ Condition Positive | Specificity = Σ True Negative / Σ Condition Negative | |

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

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

# Czułość i swoistość: Przykład

**Badanie przesiewowe na krew utajoną w kale (FOB) zastosowane u 2.030 osób w celu wykrycia raka jelita grubego**

| | | 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)

# Czułość i swoistość: Przykład

**Obliczenia:**

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

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

**Power = czułość = 1 − β**

**Likelihood ratio positive = czułość / (1 − swoistość) = 66.67% / (1 − 91%) = 7.4**

**Likelihood ratio negative = (1 − czułość) / swoistość = (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% | |

**Zatem przy dużej liczbie wyników fałszywie dodatnich i niewielkiej liczbie wyników fałszywie ujemnych dodatni wynik testu przesiewowego FOB sam w sobie słabo potwierdza nowotwór (PPV = 10%) i należy przeprowadzić dalsze badania; prawidłowo jednak identyfikuje 66,7% wszystkich nowotworów (czułość). Jednak jako test przesiewowy, wynik negatywny potwierdza, że pacjent nie ma raka (NPV = 99,5%), a na tym wstępnym etapie prawidłowo identyfikuje 91% osób, które nie mają raka (swoistość).**

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

# Macierz przekłamań

# Macierz przekłamań

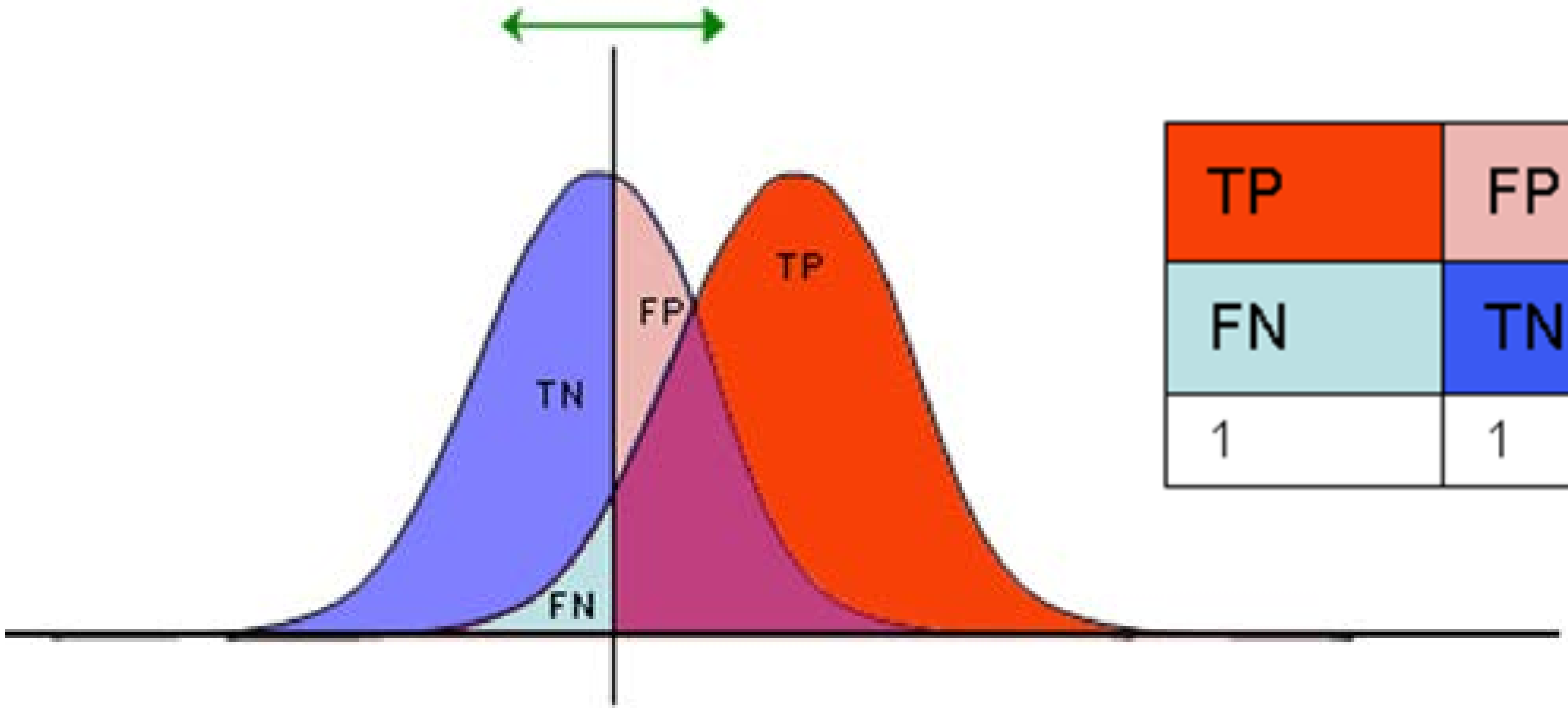
**To samo, co widzieliśmy wcześniej, ale użyte w odniesieniu do przewidywań modelu i prawdziwego stanu rzeczy.**

| | | 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)

# Krzywe ROC (Receiver Operating Characteristic)

# Krzywa ROC (interpretacja)

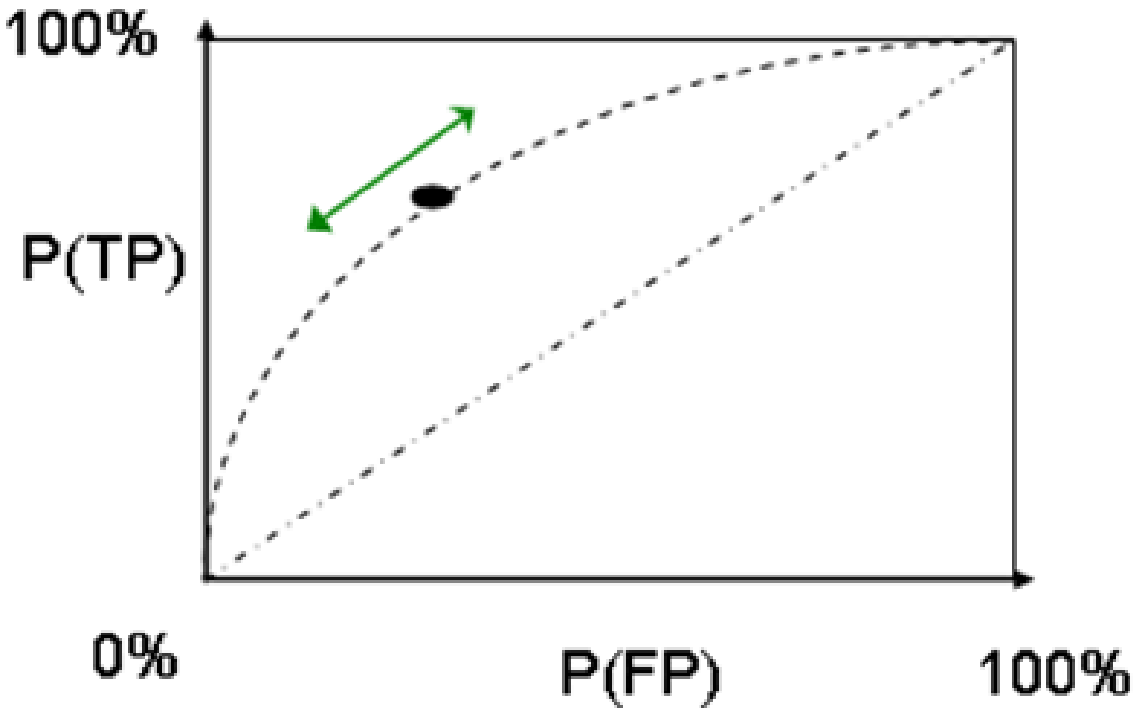



| | |
|---|---|
| TP | FP |
| FN | TN |
| 1 | 1 |

- Należy pamiętać, że bardzo często musimy znaleźć kompromis pomiędzy czułością a swoistością: wyższa czułość oznacza niższą swoistość i odwrotnie.
- Ustalenie progu jest kwestią decyzji.
- Próg, który zdecydujemy się przyjąć, określi parametry naszego testu (tj. współczynniki prawdy/fałszywie dodatniej i prawdy/fałszywie ujemnej).

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

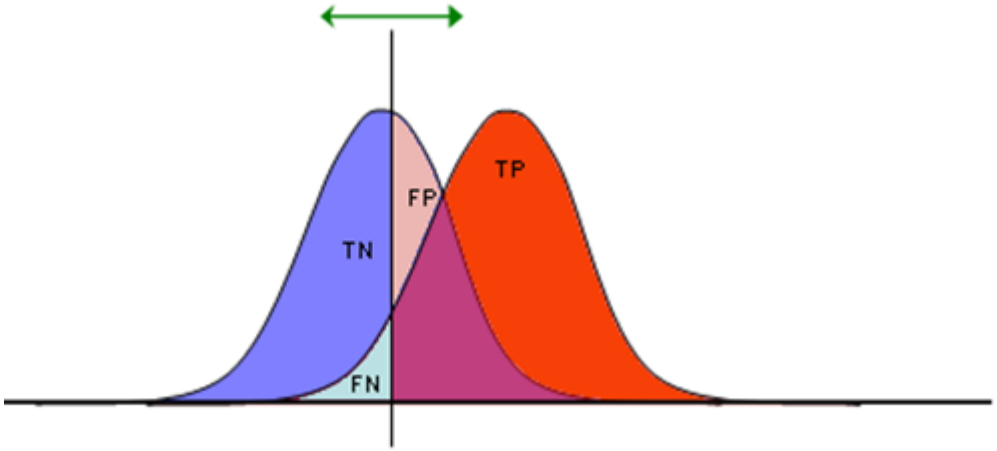
# Krzywa ROC (interpretacja)




- **Przesuwając próg, zmieniamy wartości czułości i swoistości. Wykres wszystkich możliwych wartości tych dwóch parametrów daje nam interesującą charakterystykę testu (system klasyfikacji, odbiornik itp.)**

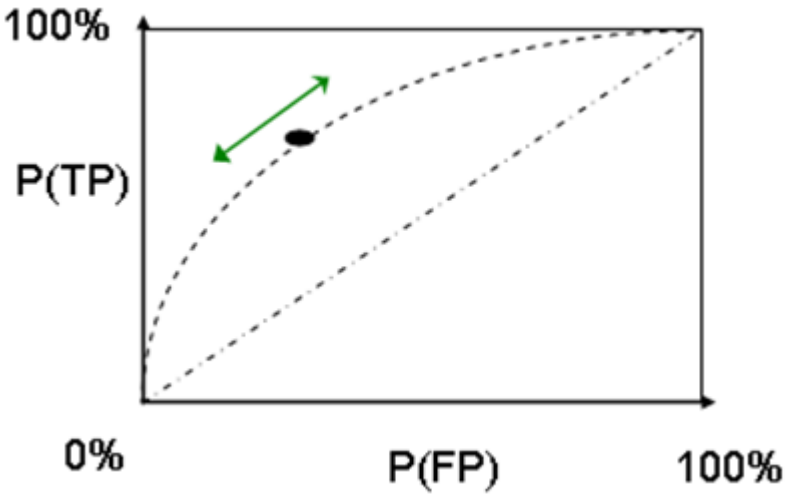
http://en.wikipedia.org/wiki/Receiver_operating_characteristic

# Krzywa ROC (interpretacja)




| | |
|---|---|
| TP | FP |
| FN | TN |
| 1 | 1 |

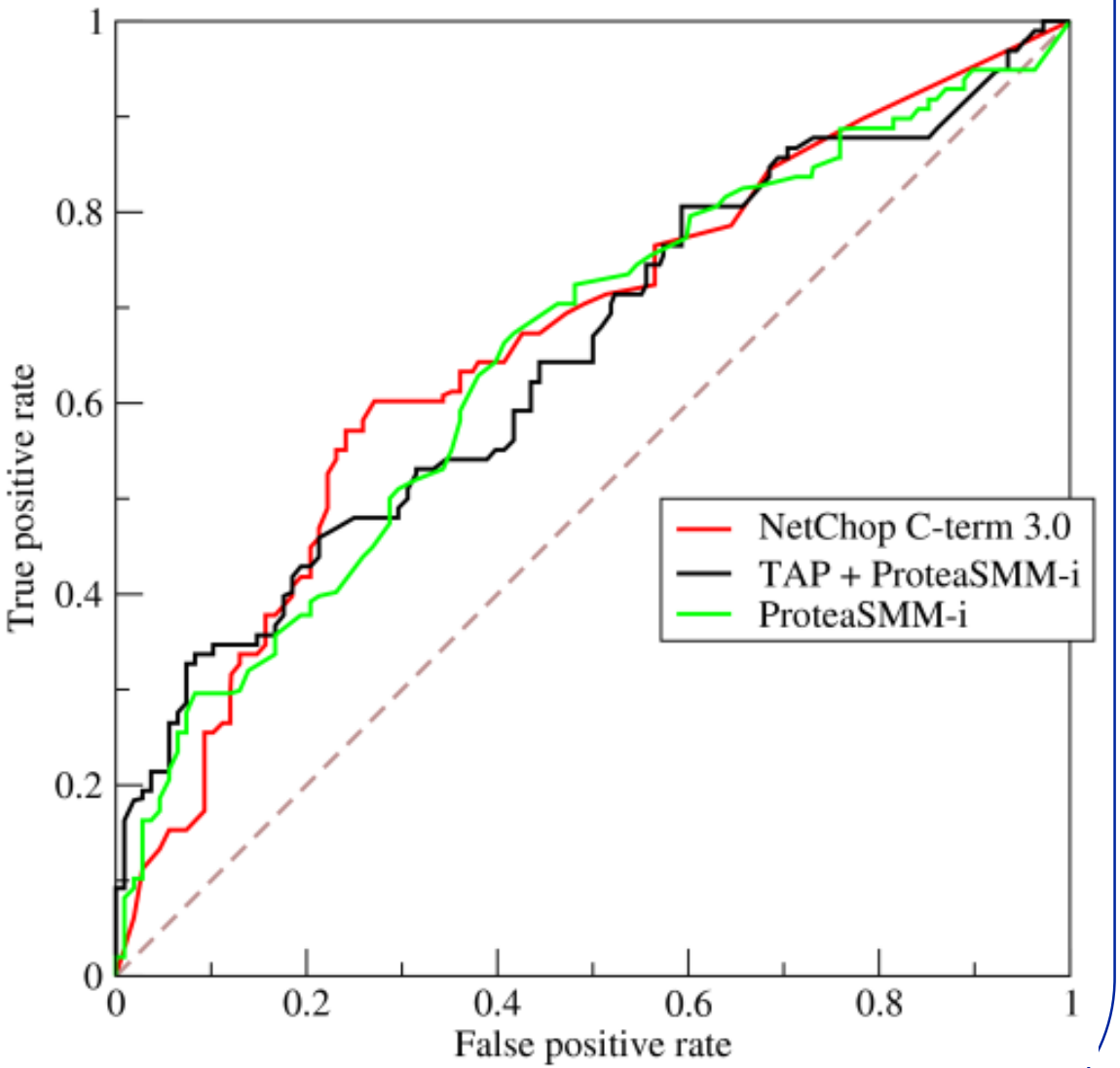
- **Wykresy takie jak ten po prawej stronie nazywane są krzywymi ROC (Receiver Operating Characteristics, charakterystyka operacyjna odbiornika).**
- **Są sposobem na scharakteryzowanie jakości systemu detekcji.**




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

# Krzywa ROC

- Wywodzi się z teorii informacji
- Jest to wykres graficzny ilustrujący działanie binarnego systemu klasyfikatorów przy zmianie jego progu dyskryminacji.

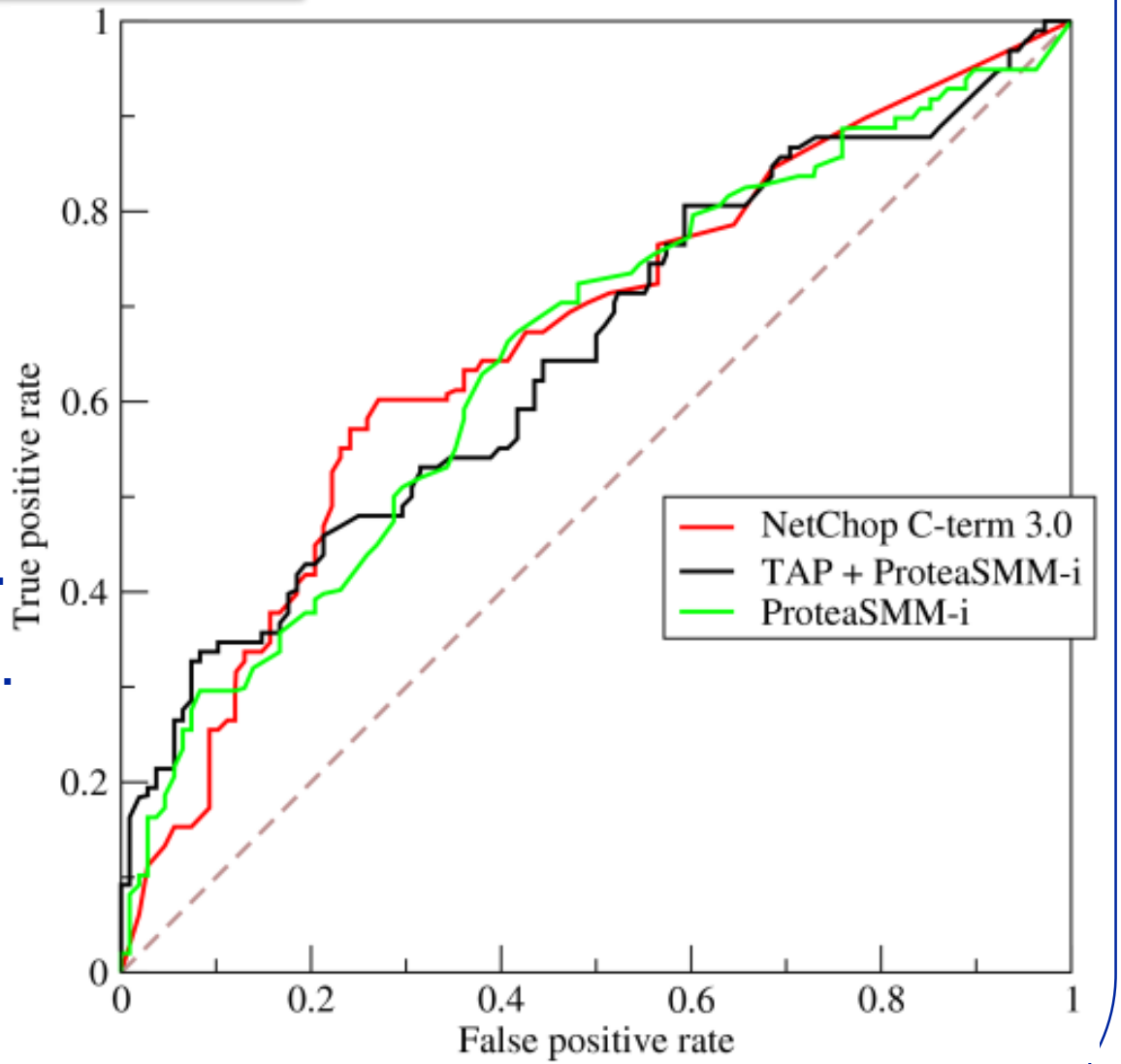



http://en.wikipedia.org/wiki/Receiver_operating_characteristic

# AUC: Area Under the (ROC) Curve (pole pod krzywą ROC)

# AUC: Pole pod krzywą ROC)

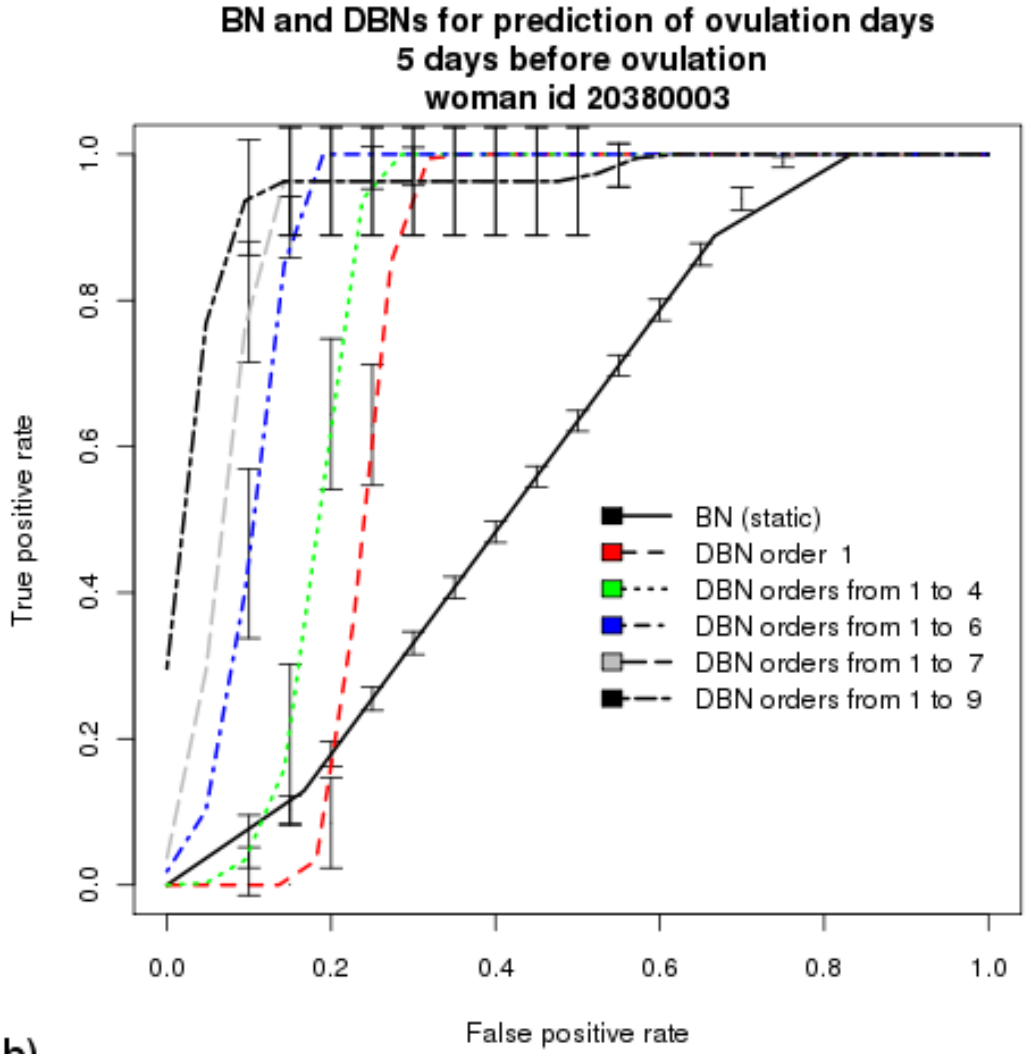
- AUC to sposób na scharakteryzowanie „odbiorcy” za pomocą jednej liczby.
- Oddaje intuicyjną koncepcję, że im wyższy ROC, tym lepiej.
- Idealna krzywa ROC przejdzie przez punkt (0,1). Pole pod taką idealną krzywą będzie wynosić 1,0.




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

# AUC: Pole pod krzywą ROC)

**AUC nie zawsze wskazuje najlepszy model**




b)

# Kalibracja

# Elementy teorii decyzji

**Teoretycznie rozsądny sposób podejmowania decyzji w warunkach niepewności**

- **Musimy wziąć pod uwagę niepewność i preferencje. Są one mierzone odpowiednio przez prawdopodobieństwo i użyteczność.**
- **Prawdopodobieństwo jest miarą niepewności.**
- **Użyteczność jest miarą preferencji, która łączy się z prawdopodobieństwem poprzez wartość oczekiwaną.**

# Przykładowa decyzja


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

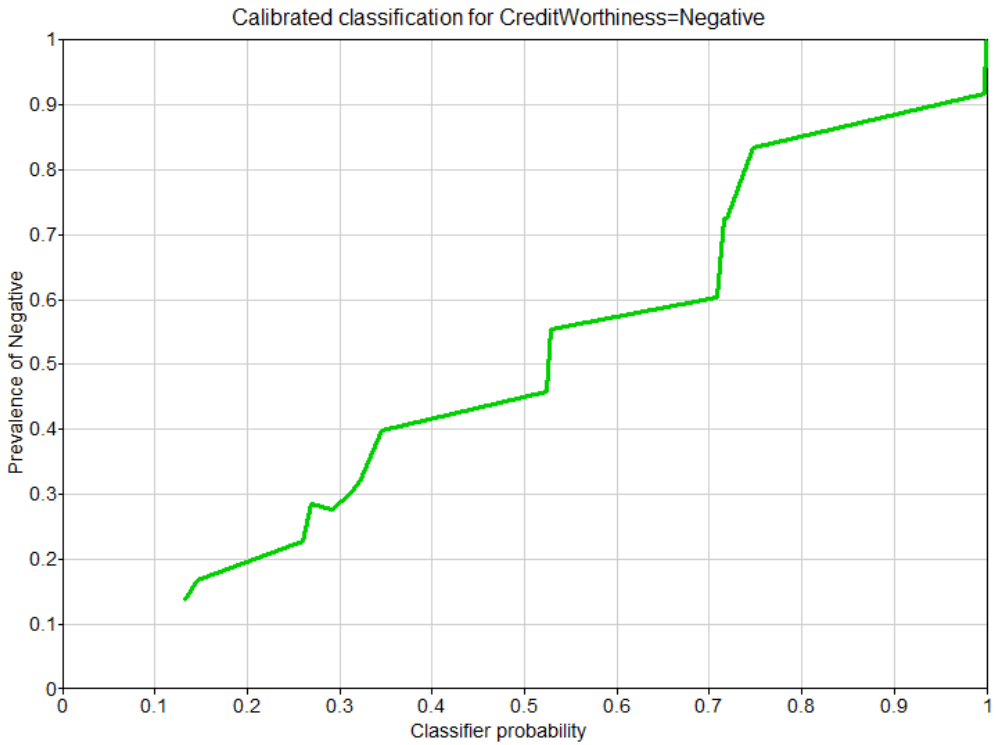
- Czy powinniśmy zabrać z domu parasol?
- Kiedy prognoza jest dobra?

# Kalibracja

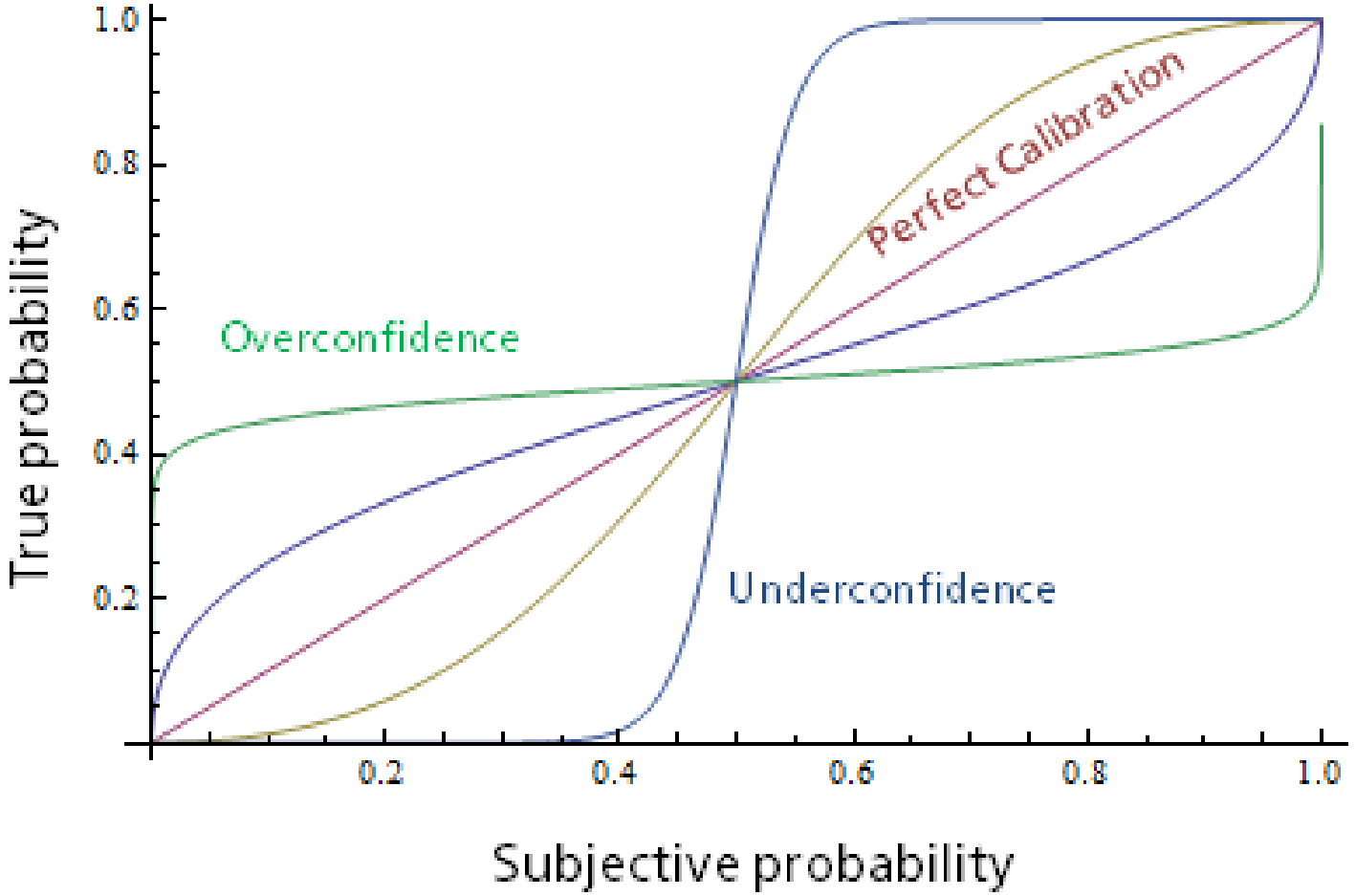
**Pytanie brzmi: czy mój model podaje dokładne prawdopodobieństwa?**

**Nanosimy częstości zaobserwowane w danych (oś rzędnych) na prawdopodobieństwa obliczone przez system (oś odciętych).**

**Różne triki wygładzające krzywą kalibracji.**

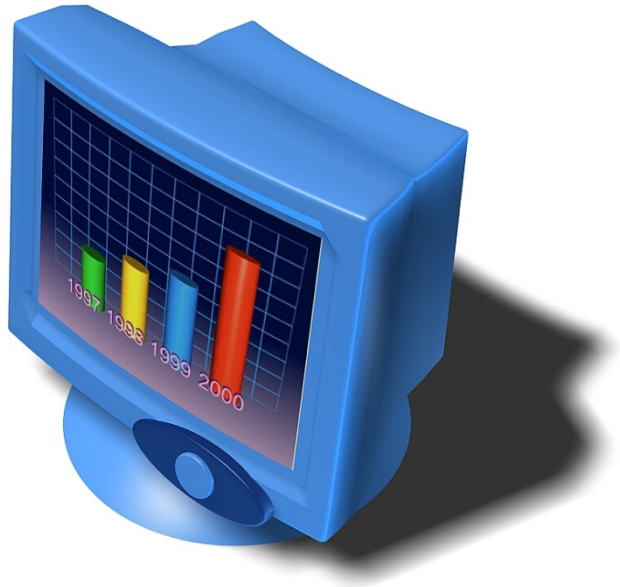


# Kalibracja: Nadmierna i niedostateczna pewność siebie




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

# Pozostałość tej sesji



# Uwagi końcowe

- Rzeczywistość jest świetnym sprawdzianem każdej aktywności ☺.
- Weryfikacja ma kluczowe znaczenie w przypadku każdego modelu i teorii, w tym modeli i teorii wywodzących się z danych.
- Statystyka ponownie jest w tym względzie światłem przewodnim.
- Istnieje wiele podejść do weryfikacji i testowania, przy czym w przypadku analizy opartej na danych dominującą metodą jest walidacja krzyżowa.
- Kiedy model generuje prawdopodobieństwo, kalibracja jest często zapominana/przeoczana.