Performance Evaluation

Hyerim Bae
Department of Industrial Engineering, Pusan National University
hrbae@pusan.ac.kr

코로나 19 검사

Source: http://news.kbs.co.kr/news/view.do?ncd=5008172&ref=A

이런 이유로 **항원·항체 검사는 유전자 검사에 비해 정확도가 떨어지는 걸로 보고**됐습니다. 식약처는 그런 점을 고려해 진단시약의 허가 기준을 항원·항체 검사의 경우 임상적 민감도 70% 이상, 특이도 90% 이상을 충족하도록 했습니다. 민감도란 질병이 있는 사람을 질병이 있다고 진단할 확률을 뜻하고 특이도는 그 반대의 경우를 말합니다. 민감도 90%, 특이도 95% 이상인 유전자 검사의 승인 기준보다 낮은 수치입니다.

보건 당국은 현재로선 이런 장점보다 **진단의 정확성이 가장 중요하다는 입장**입니다.

구분	유전자 검사	항원 검사	항체 검사
검사 목적	코로나19 바이러스 유전자 유무 확인	코로나19 바이러스 특정 단백질 유무 확인	코로나19 바이러스에 대한 항체 생성여부 확인
검사 물질	바이러스 유전자	바이러스 특정 단백질	체내 생성 항체
사용 검체	코 또는 목의 점액, 가래(객담)	코 또는 목의 점액	혈액
검사 시간	약 3 ~ 6시간	약 15분	약 15분
장점	정확도가 높아 확진용으로 사용	유전자 검사 대비 검사시간 짧고 비용 낮음	과거 감염이력 확인 가능, 검사시간 짧고 비용 낮음
단점	과거 감염 이력 확인 불가, 검사시간 길고 비용 높음	유전자 검사 대비 낮은 정확도, 확진용으로 사용 어려움	감염 초기 항체가 확인되지 않을 수 있고 검사당시 검체 내 바이러스 유무 직접 확인 어려움
측정 원리	바이러스 유전자를 증폭하여 감염여부 확인	바이러스와 결합한 특정 물질을 검출하여 바이러스 감염여부 확인	체내에 생성된 항체와 결합한 물질을 분석하여 항체 존재여부 확인
검사자 (사용자)	의료인 또는 검사 전문가	의료인 또는 검사 전문가	의료인 또는 검사 전문가

식품의약품안전처 자료.

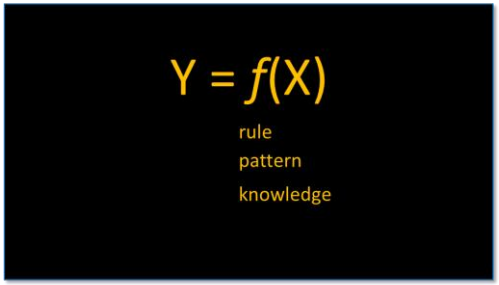
Contents

Performance evaluation Overview

Why Evaluate?

- Multiple methods are available to classify or predict
  - 인공신경망을 쓸까? 의사결정 나무를 쓸까?
- For each method, multiple choices are available for settings
  - Activation function을 ReLu를 쓸까? Sigmoid를 쓸까?
- To choose best model, need to assess each model's performance
  - 기계학습의 정의를 다시한번 생각해보자.

▪ Finding '*f*' such that



▪ We use X and Y to find '*f*'

Types of output

- Numerical value
  - 이번학기 산업데이터과학 중간고사 점수는?
- Class
  - 이번학기 산업데이터과학의 평점은?
- Tendency: probability of being a class
  - 내가 이번학기에 산업데이터 과학에서 A+를 받을 확률은?

Misclassification error

- Error = classifying a record as belonging to one class when it belongs to another class.
- Error rate = percent of misclassified records out of the total records in the validation data

Benchmark

- Benchmark: Classification Benchmark (Naïve rule): classify all records as belonging to the most prevalent class

  "가장 일반적인 클래스에 속한다고 분류"

  - Often used as benchmark: we hope to do better than that
  - Exception: when goal is to identify high-value but rare outcomes, we may do well by doing worse than the naïve rule (see "lift" – later)

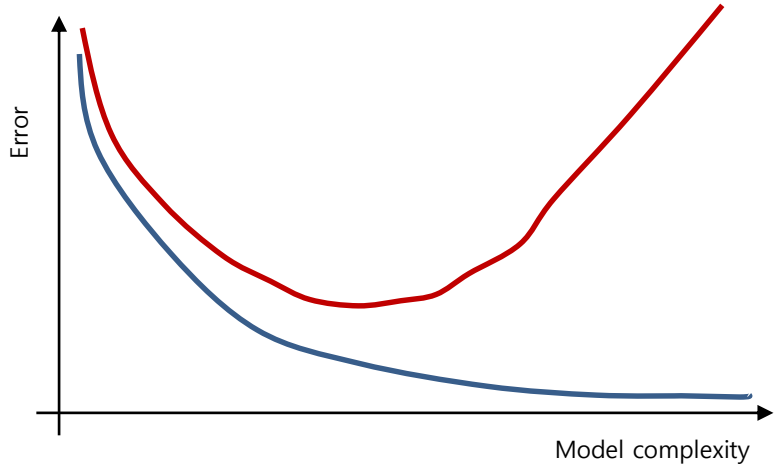
- Prediction Benchmark: Mean "학습데이터들의 평균값을 예측값으로 사용"

Error measure for prediction

- Error란: 예측값과 실제값의 차이
  - $e_i = \hat{y}_i - y_i$
- Mean Absolute Error: MAE (or MAD)
  - $1/n \sum_{i=1}^{n} |e_i|$
- Average Error: AE
  - $1/n \sum_{i=1}^{n} e_i$
- Mean Absolute Percentage Error: MAPE(평균절대 백분율 오차)
  - $100 \times 1/n \sum_{i=1}^{n} |e_i/y_i|$
- Rooted Mean Squared Error: RMSE
  - $\sqrt{1/n \sum_{i=1}^{n} {e_i}^2}$
- Sum of Squared Error: SSE
  - $\sum_{i=1}^{n} {e_i}^2$

Training vs. Validation

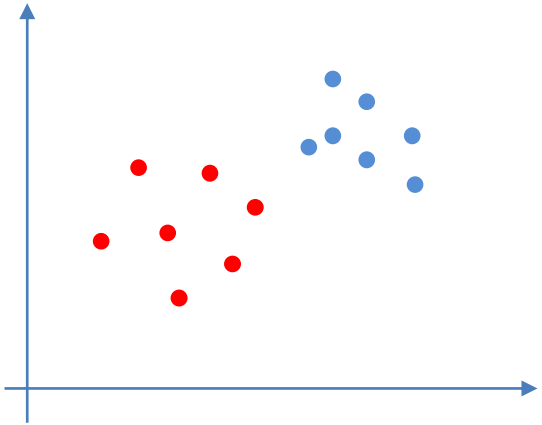
- In supervised learning
  - Training error
  - Validation error



Separation of Records in Classification

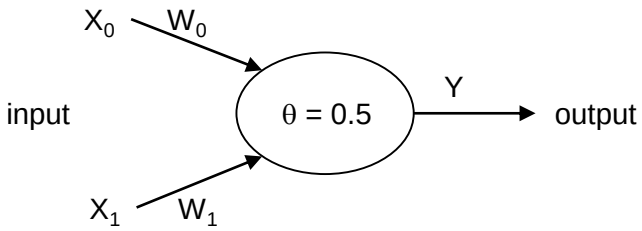
"High separation of records" means that using predictor variables attains low error

"Low separation of records" means that using predictor variables does not improve much on naïve rule



Linearly Separable

- Are AND, XOR 'linearly separable'?



input		Output (by *f*)		
X_0	X_1	AND	OR	XOR
0	0	0	0	0
0	1	0	1	1
1	0	0	1	1
1	1	1	1	0

- AND

$$
\begin{aligned}
0\times W_0 + 0\times W_1 = 0 &\quad < 0.5\\
0\times W_0 + 1\times W_1 = W_1 &\quad < 0.5\\
1\times W_0 + 0\times W_1 = W_0 &\quad < 0.5\\
1\times W_0 + 1\times W_1 = W_0 + W_1 &\quad > 0.5
\end{aligned}
$$

→ W_0, W_1 : 0.3 or 0.4

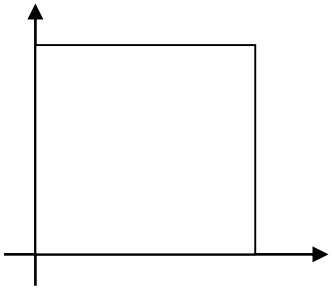
- XOR

$$
\begin{aligned}
0 \times W_0 + 0 \times W_1 = 0 &\quad < 0.5 \\
0 \times W_0 + 1 \times W_1 = W_1 &\quad > 0.5 \\
1 \times W_0 + 0 \times W_1 = W_0 &\quad > 0.5 \\
1 \times W_0 + 1 \times W_1 = W_0 + W_1 &\quad < 0.5
\end{aligned}
$$

→ W_0, W_1 do not exist that satisfy above

→ cannot solve XOR

XOR function



Confusion Matrix

201 1's correctly classified as "1" ($n_{1,1}$)

85 1's incorrectly classified as "0" ($n_{1,0}$)

25 0's incorrectly classified as "1" ($n_{0,1}$)

2689 0's correctly classified as "0" ($n_{0,0}$)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	201	85
0	25	2689

Error Rate

Overall error rate = (25+85)/3000 = 3.67%

Accuracy = 1 – err = (201+2689) = 96.33%

If multiple classes, error rate is:

(sum of misclassified records)/(total records)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	201	85
0	25	2689

Error rate, err = $\frac{(n_{0,1}+n_{1,0})}{n}$, accuracy = 1-err

Cutoff for classification

Most DM algorithms classify via a 2-step process:

For each record,

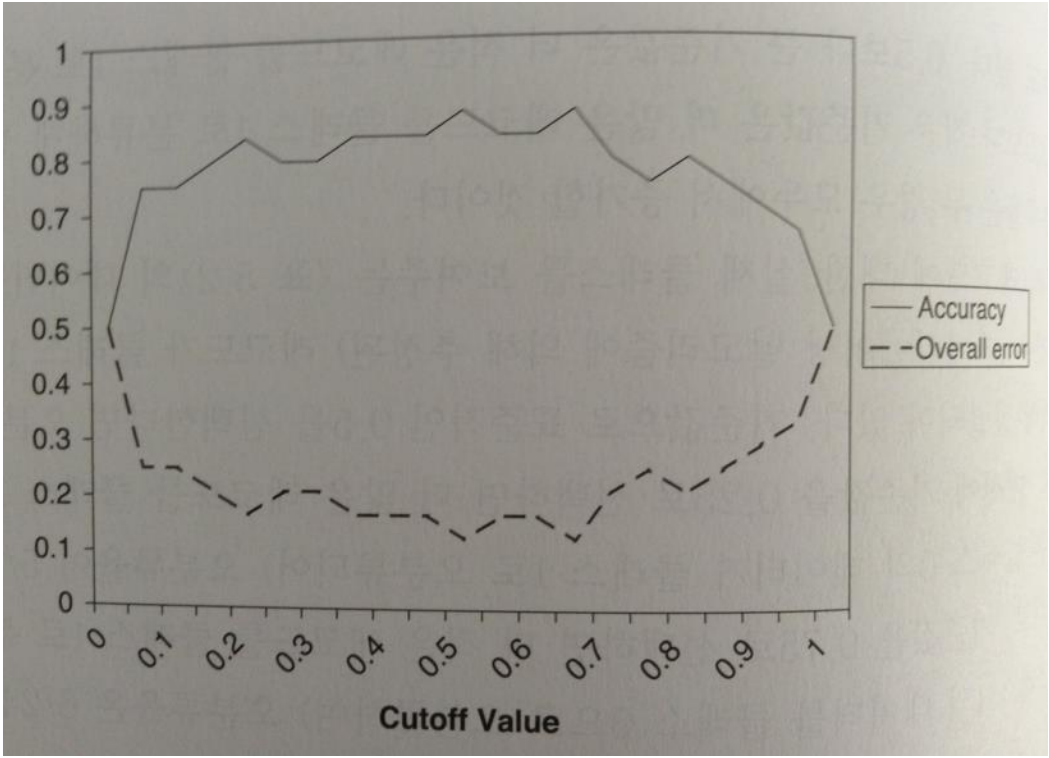
1. Compute **probability of belonging to class "1"**
2. Compare to cutoff value, and classify accordingly

- Default cutoff value is 0.50
  - If >= 0.50, classify as "1"
  - If < 0.50, classify as "0"
- Can use different cutoff values
- Typically, error rate is lowest for cutoff = 0.50

Cutoff Table

- If cutoff is 0.50: eleven records are classified as “1” (error rate = ?)
- If cutoff is 0.80: seven records are classified as “1”

Actual Class	Prob. of "1"	Actual Class	Prob. of "1"
1	0.996	1	0.506
1	0.988	0	0.471
1	0.984	0	0.337
1	0.980	1	0.218
1	0.948	0	0.199
1	0.889	0	0.149
1	0.848	0	0.048
0	0.762	0	0.038
1	0.707	0	0.025
1	0.681	0	0.022
1	0.656	0	0.016
0	0.622	0	0.004


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Cutoff Value
Accuracy
Overall error

Lift

When One Class is More Important

In many cases it is more important to identify members of one class

"회사가 파산할지를 예측하는 것이 지불능력을 유지할 지를 예측하는 것보다 더 중요하다."

- Tax fraud
- Credit default
- Response to promotional offer
- Detecting electronic network intrusion
- Predicting delayed flights

In such cases, we are willing to tolerate greater overall error, in return for better identifying the important class for further attention

Alternate Accuracy Measures

If "C_1" is the important class,

Sensitivity(민감도) = % of "C_1" class correctly classified

$$\frac{n_{1,1}}{(n_{1,0} + n_{1,1})}$$

Specificity (특이도) = % of "C_0" class correctly classified

$$\frac{n_{0,0}}{(n_{0,0} + n_{0,1})}$$

False positive rate = % of predicted "C_1's" that were not "C_1's"

$$\frac{n_{0,1}}{(n_{0,0} + n_{0,1})}$$

False negative rate = % of predicted "C_0's" that were not "C_0's"

$$\frac{n_{1,0}}{(n_{1,0} + n_{1,1})}$$

Precision and Recall

- Precision
  - In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the query:

		True condition	
		True	False
Predicted condition	True	True Positive	False Positive
	False	False Negative	True Negative

- Recall
  - the fraction of the relevant documents

that are successfully retrieved.

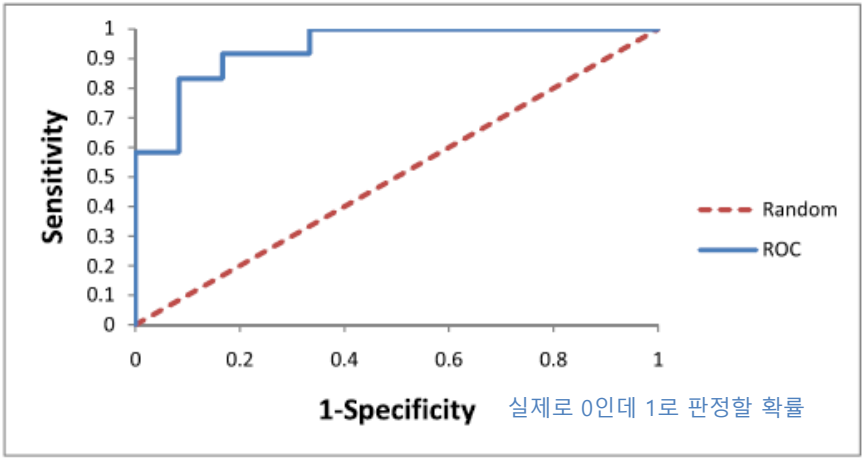
- F1 score
  - Harmonic mean of precision and recall

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

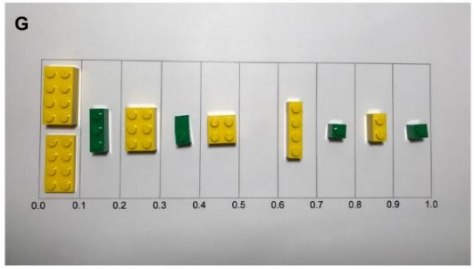
ROC Curve

- Random하게 뽑으면?

실제로 1인데 1로 판정할 확률

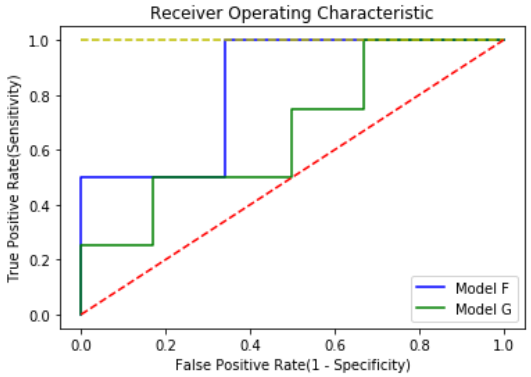


실제로 0인데 1로 판정할 확률



홀수 블록 임계값	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
맞춘 홀수(전체4개)	4	4	4	4	3	2	2	2	2	2	0
맞춘 짝수(전체6개)	0	1	3	4	4	4	5	6	6	6	6
정확도	40%	50%	70%	80%	70%	60%	70%	80%	80%	80%	60%
민감도	100%	100%	100%	100%	75%	50%	50%	50%	50%	50%	0%
특이도	0%	16.6%	50%	66.6%	66.6%	66.6%	83.3%	100%	100%	100%	100%

홀수 블록 임계값	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
맞춘 홀수(전체4개)	4	4	3	3	2	2	2	2	1	1	0
맞춘 짝수(전체6개)	0	2	2	3	3	4	4	5	5	6	6
정확도	40%	60%	50%	60%	50%	60%	60%	70%	60%	70%	60%
민감도	100%	100%	75%	75%	50%	50%	50%	50%	25%	25%	0%
특이도	0%	33.3%	33.3%	50%	50%	66.6%	66.6%	83.3%	83.3%	100%	100%



Lift and Decile Charts

Useful for assessing performance in terms of identifying the most important class

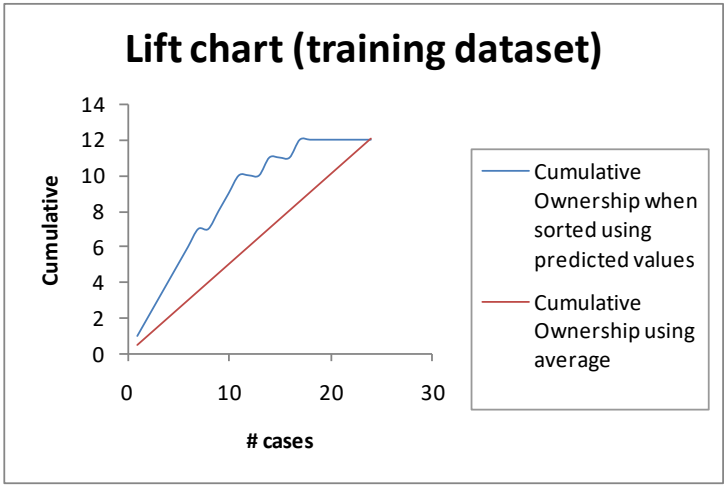
Helps evaluate, e.g.,

- How many tax records to examine
- How many loans to grant
- How many customers to mail offer to

Lift Chart – cumulative performance

After examining (e.g.,) 10 cases (x-axis), 9 owners (y-axis) have been correctly identified

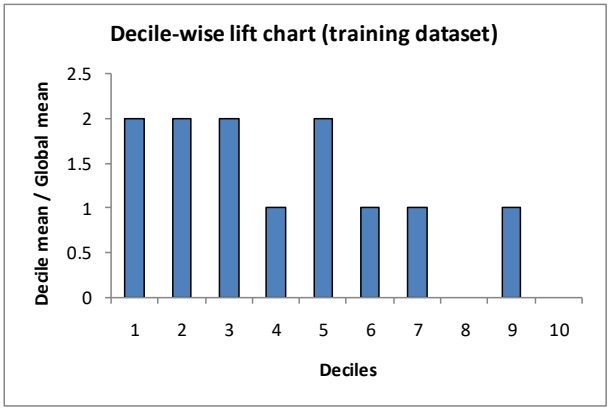
좋은 분류기는 적은수의 케이스만으로 높은향상을 제공



Actual Class	Prob. of "1"	Actual Class	Prob. of "1"
1	0.996	1	0.506
1	0.988	0	0.471
1	0.984	0	0.337
1	0.980	1	0.218
1	0.948	0	0.199
1	0.889	0	0.149
1	0.848	0	0.048
0	0.762	0	0.038
1	0.707	0	0.025
1	0.681	0	0.022
1	0.656	0	0.016
0	0.622	0	0.004

Decile Chart

In “most probable” (top) decile, model is twice as likely to identify the important class (compared to avg. prevalence)



Lift vs. Decile Charts

Both embody concept of “moving down” through the records, starting with the most probable

Decile chart does this in decile chunks of data

- Y axis shows ratio of decile mean to overall mean

Lift chart shows continuous cumulative results

- Y axis shows number of important class records identified

Asymmetric Costs

Misclassification Costs May Differ

The cost of making a misclassification error may be higher for one class than the other(s)

Looked at another way, the benefit of making a correct classification may be higher for one class than the other(s)

Example – Response to Promotional Offer

Suppose we send an offer to 1000 people, with 1% average response rate ("1" = response, "0" = nonresponse)

- "Naïve rule" (classify everyone as "0") has error rate of 1% (seems good)
- Using DM we can correctly classify eight 1's as 1's
  It comes at the cost of misclassifying twenty 0's as 1's and two 0's as 1's.

The Confusion Matrix

Error rate = (2+20) = 2.2% (higher than naïve rate)

	Predict as 1	Predict as 0
Actual 1	8	2
Actual 0	20	970

Introducing Costs & Benefits

Suppose:

- Profit from a "1" is $10
- Cost of sending offer is $1

Then:

- Under naïve rule, all are classified as "0", so no offers are sent: no cost, no profit
- Under DM predictions, 28 offers are sent.

  8 respond with profit of $10 each
  20 fail to respond, cost $1 each
  972 receive nothing (no cost, no profit)

- Net profit = $60

Profit Matrix

	Predict as 1	Predict as 0
Actual 1	$80	0
Actual 0	($20)	0

Lift (again)

Adding costs to the mix, as above, does not change the actual classifications

Better: Use the lift curve and change the cutoff value for "1" to maximize profit

Note: Opportunity costs

- As we see, best to convert everything to costs, as opposed to a mix of costs and benefits
- E.g., instead of “benefit from sale” refer to “opportunity cost of lost sale”
- Leads to same decisions, but referring only to costs allows greater applicability

Adding Cost/Benefit to Lift Curve

- Sort records in descending probability of success
- For each case, record cost/benefit of actual outcome
- Also record cumulative cost/benefit
- Plot all records

  X-axis is index number (1 for 1st case, n for nth case)

  Y-axis is cumulative cost/benefit

  Reference line from origin to y_n (y_n = total net benefit)

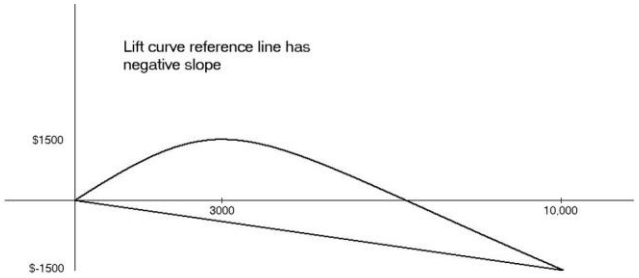
Lift Curve May Go Negative

If total net benefit from all cases is negative, reference line will have **negative slope**

Nonetheless, goal is still to use cutoff to select the point where net benefit is at a maximum

Negative slope to reference curve

Cost for sending one mail 0.65$, befit from the respondent 25$, response rate 2%,
* If we send to 10,000 people? (0.02*$25*10,000) – (0.65*10,000) = - 1500



Summary

- Model evaluation
  - Evaluation metrics are important for comparing across DM models, for choosing the right configuration of a specific DM model, and for comparing to the baseline
  - Major metrics: confusion matrix, error rate, predictive error
  - Other metrics when
    - one class is more important
    - asymmetric costs