Applied Machine Learning
Lecture 6-2: Evaluation methods

Selpi (selpi@chalmers.se)

The slides are further development of Richard Johansson's slides

February 11, 2020

Overview

why evaluate?

- predictive systems are rarely perfect
- so we need to **measure** how well our systems are doing
  - and compare alternatives
- selecting a meaningful evaluation protocol is crucial in machine learning projects
  - this is something we emphasize for thesis projects
- what is the performance needed to be "useful"?

how do we evaluate? intrinsic and extrinsic

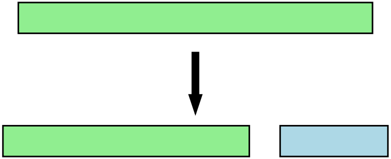
- **intrinsic** evaluation: measure the performance in isolation, using some automatically computed metric
- **extrinsic** evaluation: I changed my predictor – how does this affect the performance of my "downstream task"?
  - how much more money do I make?
  - how many more clicks do I get?

is there already an established evaluation procedure?

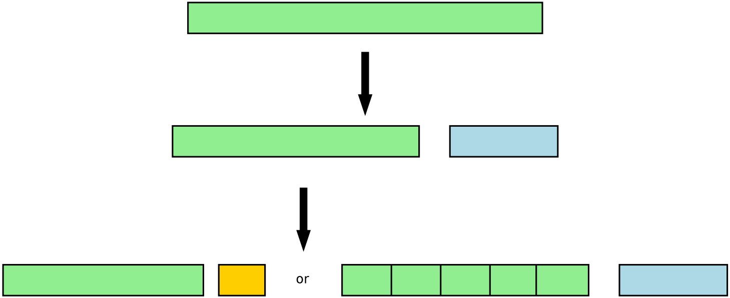
- for many classical prediction problems, there might be a well-established procedure
- some of which can be quite specific to the type of application:
  - word error rate for speech recognition systems
  - BLEU score for machine translation systems
- in this lecture, we'll look at some of the most common **generally applicable** evaluation methods that are
  - intrinsic
  - automatic
- we'll come back to some application-specific metrics at the end

high-level setup

high-level setup



high-level setup



qualitative analysis

- it can be useful to look at "interesting" instances
  - but not as an argument for why your system is good
  - or for systematic comparison
- error analysis – an art more than a science
  - we need to understand our data
- can help me understand what data I lack, or help me improve features

Overview

accuracy and error rate

$$\text{accuracy} = \frac{\text{nbr. correct}}{N}$$

$$\text{error rate} = \frac{\text{nbr. incorrect}}{N}$$

example

confusion matrix

		Actual class		
		Cat	Dog	Rabbit
Predicted class	Cat	5	2	0
	Dog	3	3	2
	Rabbit	0	1	11

[source]

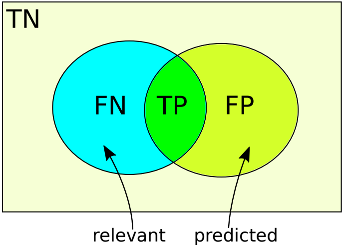
considering one class

- let's say we're interested in one class in particular
  - "spotting problems" or "finding the needles in the haystack"

		Actual class	
		Cat	**Non-cat**
Predicted class	**Cat**	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	17 True Negatives

evaluation metrics derived from a confusion matrix

- several evaluation metrics are defined from the cells in a confusion matrix
  - depending on the scientific field
- these metrics often come in pairs
  - overgeneration vs undergeneration



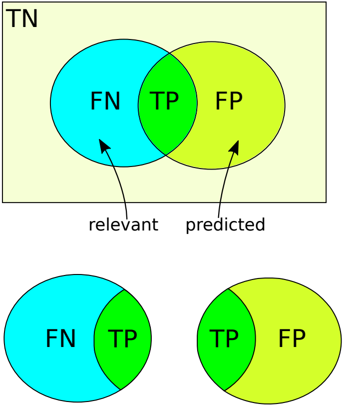
see https://en.wikipedia.org/wiki/Confusion_matrix

precision and recall

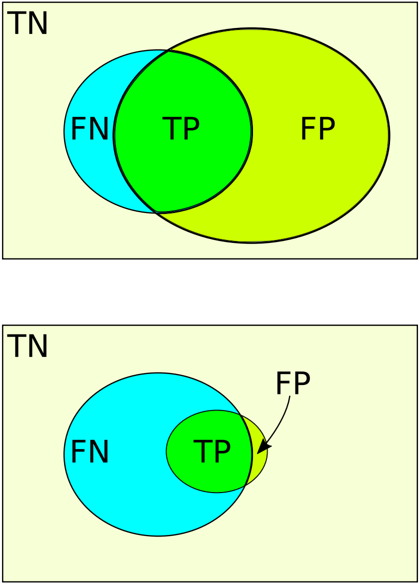
- the **precision** and **recall** are commonly used for evaluating classifiers

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN}$$

- in scikit-learn:
  - `sklearn.metrics.precision_score`
  - `sklearn.metrics.recall_score`
  - `sklearn.metrics.classification_report`



precision / recall tradeoff



F-score

- the *F*-**score** is the harmonic mean of the precision and recall

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

- in scikit-learn: `sklearn.metrics.f1_score`

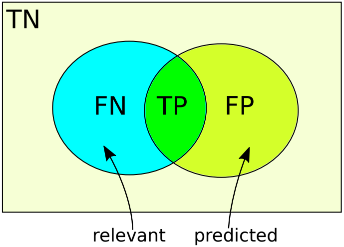
other similar pairs of metrics

- **sensitivity/specificity** (in medicine)

$$\text{Sen} = \frac{TP}{TP + FN} \qquad \text{Spe} = \frac{TN}{TN + FP}$$

- **true positive rate/false positive rate**

$$\text{TPR} = \frac{TP}{TP + FN} \qquad \text{FPR} = \frac{FP}{FP + TN}$$



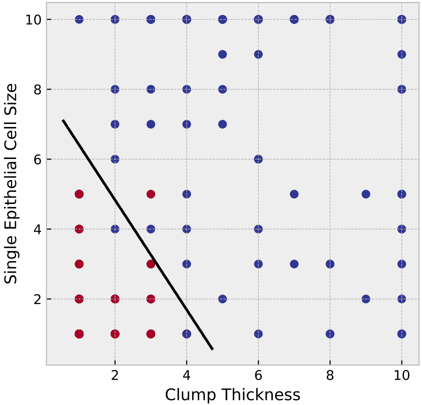
example (continued)

Overview

what if it's more important to find the malignant tumors?

- how can we adjust our decisions?

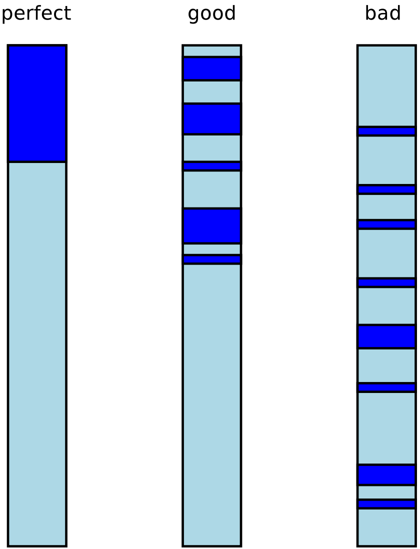
breast cancer example



ranking and scoring systems

- for instance, search engines
- but also some classifiers (e.g. `decision_function`, `predict_proba`)

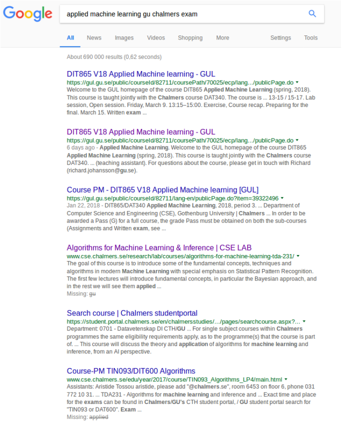
intuition: evaluating rankers



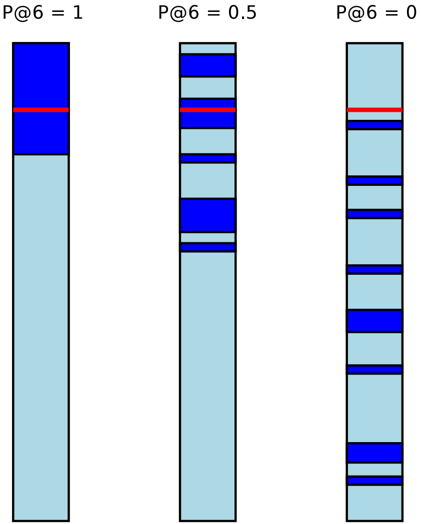
evaluating a search engine: quality of the first page

- when I use a search engine, how good are the results on the first page?
- **precision at k**

$$\text{P@}k = \frac{\text{nbr relevant items in top } k}{k}$$



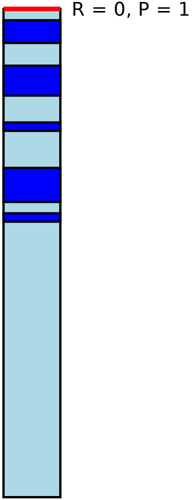
evaluating a search engine (another example)

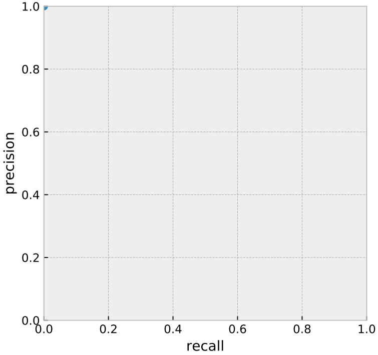


precision / recall curve

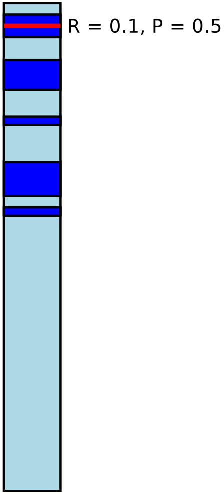
- is there a way to evaluate ranker or scorer that does not depend on the threshold?
- in a **precision/recall curve**, we compute the precision for different recall levels
  - that is, for different thresholds
- we visualize the relationship using a plot
  - good = curve is near top right

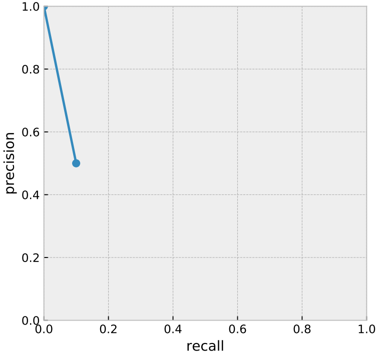
P / R curve (example)



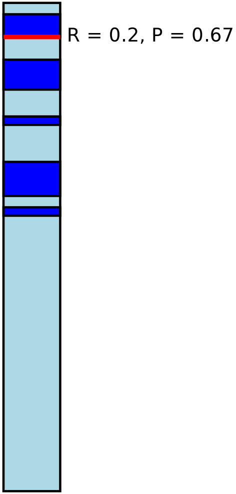


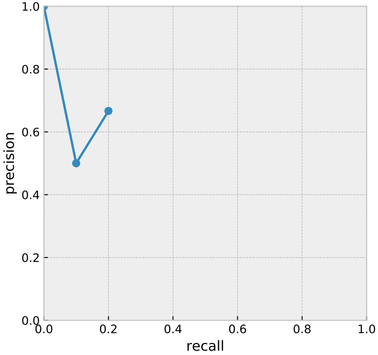
P / R curve (example)



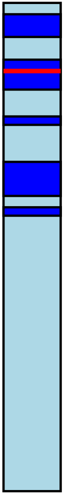


P / R curve (example)

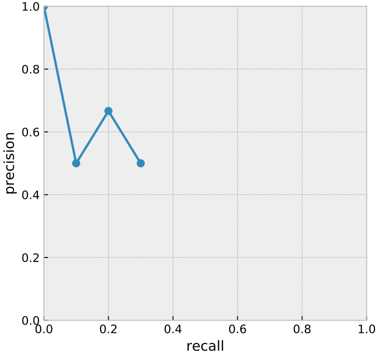




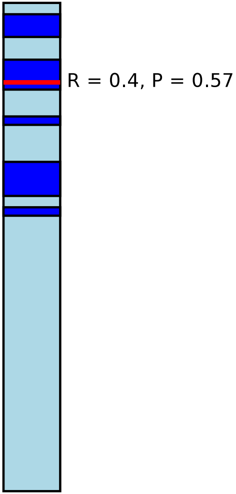
P / R curve (example)

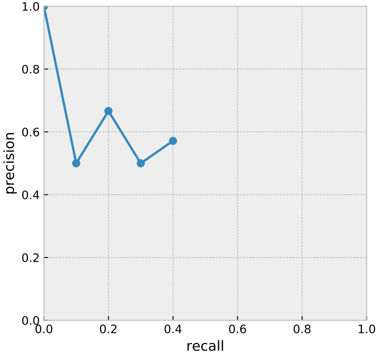


R = 0.3, P = 0.5

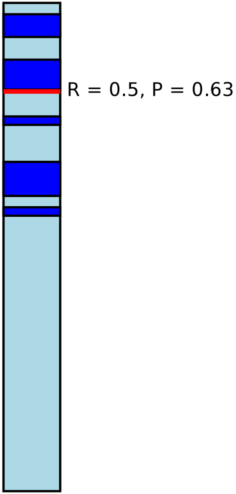


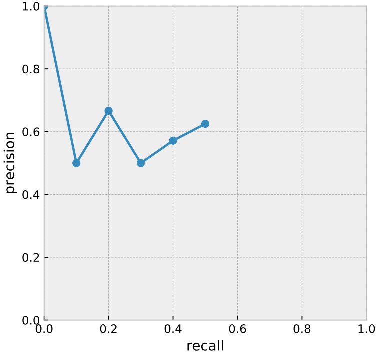
P / R curve (example)



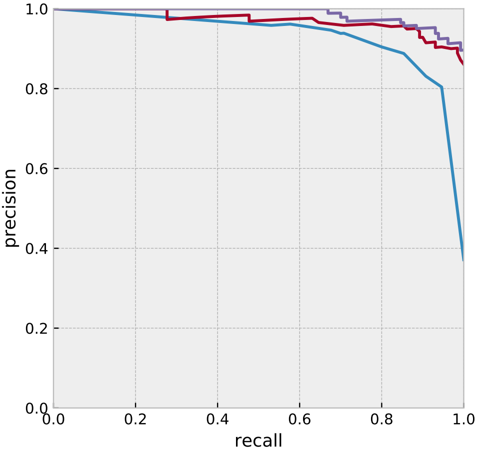


P / R curve (example)

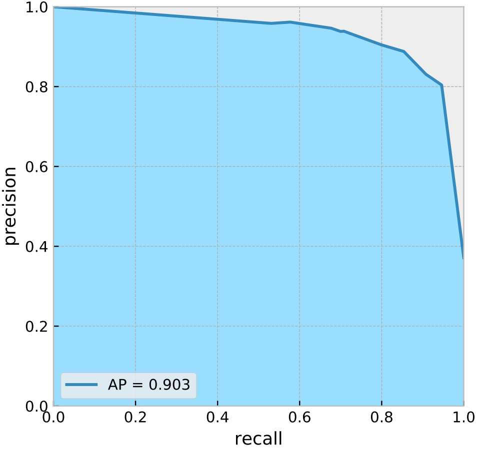




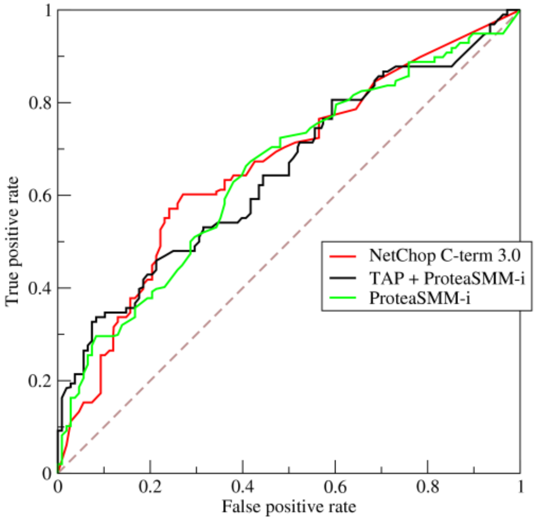
comparing P/R curves



average precision

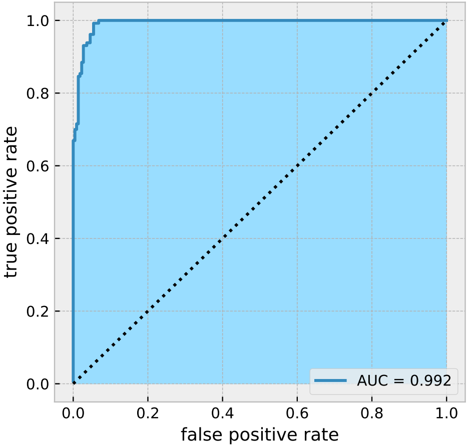


ROC curve (Receiver Operating Characteristic)



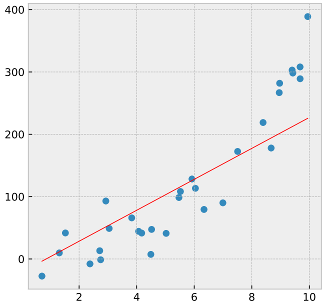
[source]

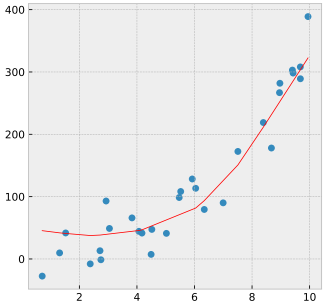
Area Under Curve score



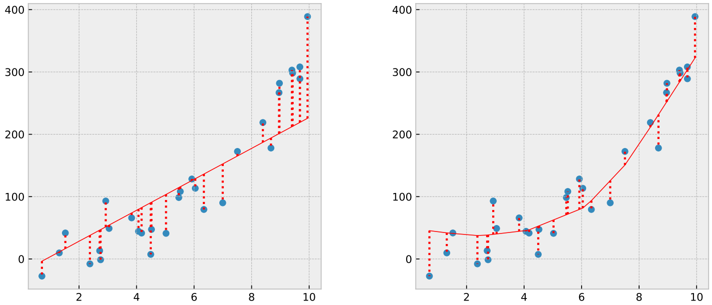
Overview

example





errors



mean squared error and mean absolute error

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad \mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

(N = number of instances, $\hat{y}_i$ = prediction for instance i)

R^2: the coefficient of determination

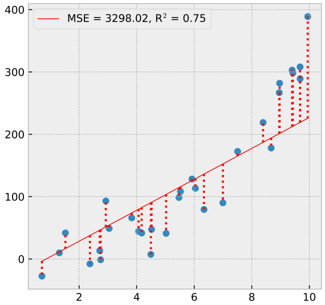
- the MSE and MAE scores depend on the scale
- they don't directly tell whether the regressor behaves "well"

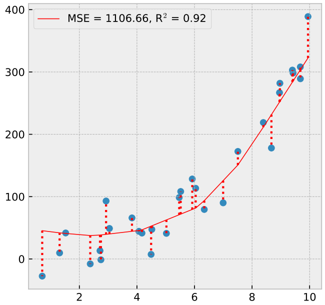
- the **coefficient of determination** (R^2) is an evaluation score that does not depend on the scale

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

- in case of a perfect regression, this score is 1
- if completely messed up, 0 or negative

MSE and R^2 for the example





what to do about ordinal scales

Grekisk pizzeria
Plaka
Utlandagatan 14

Det är inte mycket som lockar till en pizzeria i dag eftersom det är oftast är bukfylla utan vidare ambitioner, men ibland finns en parallell taffel som kan glimra i hemlighet. Vi gick till Plaka som ligger i ett annars restaurangsnålt område.

Vi anade snart att det var servitrisen som var kvällens blyga behållning – snabb, vänlig, lyhörd. Plaka har en pizzameny med hela 49 varianter mellan 85 och 95 kronor som slutar med en Gyros Pizza. Det föreföll lika äckligt som Kebab Pizza. En maträtt från helvetet som kanske borde returneras dit.

Därför sneglade vi mot den grekiska menyn med den famlande tanken om att den grekiska krisen kanske ändå inte nått ända dit. Vi beställde två standardförrätter. Grillad halloumi och friterad bläckfisk. I halvfabrikatens värld fordras inget geni direkt

- see e.g. Baccianella et al. (2009) *Evaluation Measures for Ordinal Regression*

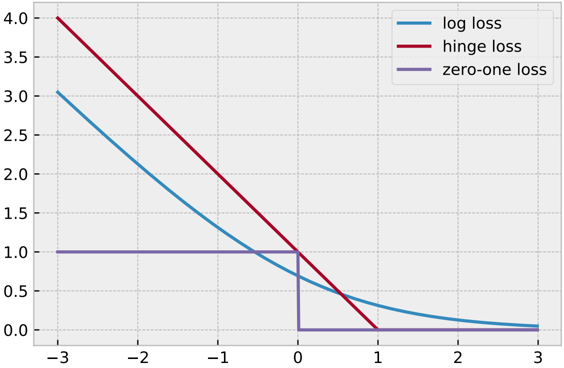
Overview

regression and classification evaluation

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad \text{error} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[y_i \neq \hat{y}_i]$$

- for **regression**, MSE is often used
  - as the **quality metric** when we evaluate
  - as the **loss function** when we train
- but typically, for **classification** what we compute during training and testing is different
  - such as hinge loss or log loss during training

loss functions



Overview

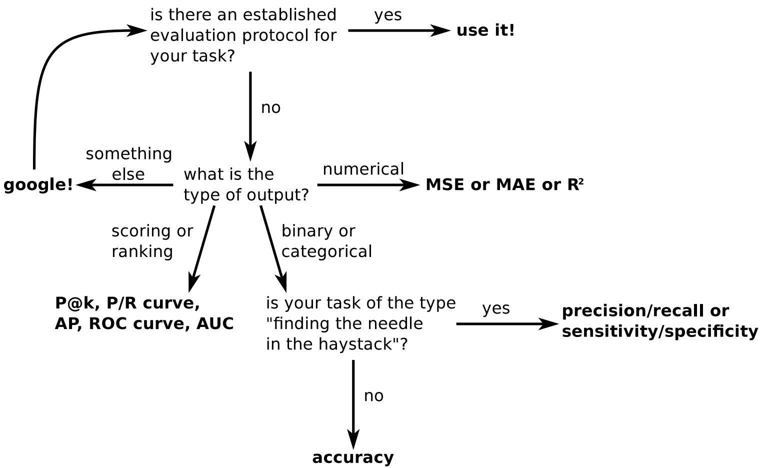
application-specific quality metrics

- as we discussed, there may be an established evaluation protocol for the task you're working on
- or you might design a new one

humans in the loop

- evaluating using **people** can be an alternative to automatic evaluation
  - if we don't have data for evaluating automatically
  - or if we want to evaluate something that is hard to measure
- expensive!
- reliable?
- replicable?

to conclude



is there an established evaluation protocol for your task?
yes
use it!
no
what is the type of output?
something else
google!
numerical
MSE or MAE or R²
scoring or ranking
P@k, P/R curve, AP, ROC curve, AUC
binary or categorical
is your task of the type "finding the needle in the haystack"?
yes
precision/recall or sensitivity/specificity
no
accuracy

Next lecture

- Neural networks