Advanced School in Artificial Intelligence

Performance Metrics

Elena Bellodi
elena.bellodi@unife.it

RegioneEmilia-Romagna

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

FERRARIAE UNIVERSITAS · EX LABORE FRUCTUS · 13 91

Università degli Studi di Ferrara

Outline

- Metrics for Hard Prediction
- Metrics for Ranking Prediction
- Metrics for Regression

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Outline

- Metrics for Hard Prediction
- Metrics for Ranking Prediction
- Metrics for Regression

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Predictive machine learning scenarios

- **Hard Prediction (Classification)**: Predict a single category for each instance
  - *Accuracy, Error, Precision, Recall (Sensitivity), Specificity, F1 Score*

- **Ranking Prediction**: learn a model that outputs a score vector $\boldsymbol{s}(x)=(s_1(x),\dots,s_k(x))$ over the k classes
  - $s_i(x)$ is the score assigned to class C_i for instance x. This score indicates how likely it is that class label C_i applies. If we only have 2 classes, $s(x)$ denotes the score of the positive class for x
  - *ROC, Precision-Recall Curves*

Predictive machine learning scenarios

- **Probability Estimation**: learn a model that outputs a probability vector over classes
- **Regression:** learn an approximation $g : X \rightarrow \mathbb{R}$ to the true labelling function f

- The metrics that you choose to evaluate your machine learning model is very important. Choice of metrics influences how the performance of machine learning algorithms is measured and compared

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for hard prediction

Confusion matrix

- For **classification problems** where the output can be of two or more classes
- **Each column refers to actual classes as recorded in the test set**
- **Each row to classes as predicted by the classifier**

Actual (columns) / **Predicted** (rows)

	Positives	Negatives
Positives	TP	FP
Negatives	FN	TN

Assume positive label as true and negative label as false

True positive (TP): predicted class true coincides with actual one, which is true

True negative (TN): predicted false class coincides with actual false class

False positive (FP): the actual data is false and the predicted is true

False negative (FN): the actual data is true while the predicted is false

Metrics for hard prediction

Confusion matrix

		Actual	
		Positives	Negatives
Predicted	Positives	TP	FP
	Negatives	FN	TN

- Find a solution that maximizes TP and TN and minimizes FP and FN
- In the best solution FN and FP are 0

Metrics for hard prediction

Confusion matrix

Actual

Predicted

	Positives	Negatives	Marginals
Positives	**30**	10	40
Negatives	20	**40**	60
Marginals	50	50	100

Actual

Predicted

	Positives	Negatives	Marginals
Positives	**20**	30	50
Negatives	20	**30**	50
Marginals	40	60	100

same marginals, but the classifier makes a random choice

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for hard prediction

Accuracy

- the proportion of **correctly classified test instances**

Predicted \ Actual	Positives	Negatives
Positives	TP	FP
Negatives	FN	TN

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Good when the number of examples for each label is nearly balanced.
Really bad when the set of examples is unbalanced (data are a majority of one class).

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for hard prediction

Accuracy

- The number of correct predictions made by the model.

Actual

Positives Negatives

Predicted

Po

Ne

If we have 100 test examples, 5 examples are + and 95 are -, even when the model predicts ALWAYS NEGATIVE the accuracy is 95%

Good when the number of examples for each [...] is nearly balanced

Really bad when the set of examples is unbalanced (data are a majority of one class)

Regione Emilia-Romagna

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Università degli Studi di Ferrara

Metrics for hard prediction

Error rate

- The proportion of incorrectly classified test instances

Actual

Predicted	Positives	Negatives
Positives	TP	FP
Negatives	FN	TN

$$Error\ rate = \frac{FP + FN}{TP + TN + FP + FN}$$

Equal to $1 - \boldsymbol{Accuracy}$

Metrics for hard prediction

Error rate

- The proportion of incorrectly classified test instances

Actual

Predicted	Positives	Negatives
Positives	TP	FP
Negatives	FN	TN

$$Error\ rate = \frac{FP + FN}{TP + TN + FP + FN}$$

Equal to $\mathbf{1 - \mathit{Accuracy}}$

Metrics for hard prediction

True Positive Rate and False Positive Rate

- The proportion of examples classified as positive (negative) among those that are actually true (false)

		Actual: Positives	Actual: Negatives
Predicted	Positives	TP	FP
	Negatives	FN	TN

$$TP\ Rate = \frac{TP}{TP + FN}$$

$$FP\ Rate = \frac{FP}{FP + TN}$$

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for hard prediction

Precision (P)

- The proportion of examples predicted as true which are actually true

		Actual	
		Positives	Negatives
Predicted	Positives	TP	FP
	Negatives	FN	TN

$$Precision = \frac{TP}{TP + FP}$$

Metrics for hard prediction

Precision (P)

- The proportion of examples predicted as true which are actually true

$$Precision = \frac{TP}{TP + FP}$$

- **Good** when we need to minimize false positives
  - If we have 100 examples, 5 examples are +, if the model predicts ALWAYS TRUE P = 5 / (5+95) = 0.05 = 5% (FP high: 95)
  - If it predicts ALWAYS FALSE except for 1 + example classified as +, P = 1 / (1+0) = 100% (FP: 0)
- **Precision** is about being precise. So even if we managed to capture only one cancer case, and we captured it correctly, then we are 100% precise.

Metrics for hard prediction

- Precision is a **counterpart** to TP rate:
  - TP rate is the proportion of predicted positives among the actual positives
  - P is the proportion of actual positives among the predicted positives

$$Precision = \frac{TP}{TP + FP}$$

- **If the minority class is the class of interest and very small**, accuracy and performance on the majority class are not the right quantities to optimise → USE PRECISION instead

Metrics for hard prediction

Recall (R) or Sensitivity

- **Equal to TP Rate**

		Actual	
		Positives	Negatives
Predicted	Positives	TP	FP
	Negatives	FN	TN

$$Recall = \frac{TP}{TP + FN}$$

If we have 100 examples, 5 examples are +, if the model predicts ALWAYS TRUE recall is 100%, ALWAYS FALSE has recall 0%, ALWAYS FALSE except for ONE correct true has recall 20%

- **Good** when we need to minimize false negatives.
- Recall is not so much about capturing cases correctly but more about capturing all cases that have “cancer” with the answer as “cancer”.

Metrics for hard prediction

- So basically if we want to focus more on **minimizing false negatives**, we would want our Recall to be as close to 100% as possible without precision being too bad
- if we want to focus on **minimizing false positives**, then our focus should be to make Precision as close to 100% as possible.

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021

Metrics for hard prediction

Specificity

- The proportion of false examples that are predicted as false.

Actual

Predicted	Positives	Negatives
Positives	TP	FP
Negatives	FN	TN

$$Specificity = \frac{TN}{TN + FP}$$

Good when we need to minimize false positive.
The exact opposite of Recall.

If we have 100 examples, 5 examples are +, if the model returns ALWAYS TRUE the specificity is 0%, ALWAYS FALSE the specificity is 100%

Metrics for hard prediction

F-Measure or F1-Score

- Considers Precision and Recall to give a score that represents both.
- Computed as the Harmonic mean

$$FMeasure = \frac{2 * Precision * Recall}{Precision + Recall}$$

With Recall=40% and Precision=60% the F-Measure is 48%.
With Recall=5% and Precision=100% the F-Measure is 9.5%.

If Accuracy and Recall are similar, the F-Measure behaves similar to an arithmetic mean, but as the difference between the two values increases, the **F-Measure returns a score that tends to follow the lowest value the higher the difference is.**

Next…

Metrics for ranking prediction

Progetto di alta formazione in ambito tecnologico economico e culturale per una regione della conoscenza europea e attrattiva approvato e cofinanziato dalla Regione Emilia-Romagna con deliberazione di Giunta regionale n. 1625/2021