



Искусственный интеллект в науках о Земле

Михаил Криницкий

К.Т.Н.,
зав. Лабораторией машинного обучения в науках о Земле МФТИ
с.н.с. Институт океанологии РАН им. П.П. Ширшова



Задачи классификации

Оценка качества моделей классификации

Михаил Криницкий

к.т.н.,
зав. Лабораторией машинного обучения в науках о Земле МФТИ
с.н.с. Институт океанологии РАН им. П.П. Ширшова

Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

Самая простая и (чаще всего) неверная мера качества в задачах классификации – доля верных ответов (accuracy)

$$Accuracy(y_{pred}, y_{true}) = \frac{\sum[y_{pred} == y_{true}]}{N}$$

Представим следующую задачу (и решение):

В наборе данных \mathcal{T} 95% объектов – класса А, остальные объекты – классов В, С, D

Мы создали и обучили модель, которая для любого нового объекта выдает результат «это объект класса А»

Каково значение меры качества *Accurасу* для такой модели?

Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

Самая простая и (чаще всего) неверная мера качества в задачах классификации – доля верных ответов (accuracy)

$$Accuracy(y_{pred}, y_{true}) = \frac{\sum[y_{pred} == y_{true}]}{N}$$

Представим следующую задачу (и решение):

В наборе данных \mathcal{T} 95% объектов – класса А, остальные объекты – классов В, С, D

Мы создали и обучили модель, которая для любого нового объекта выдает результат «это объект класса А»

Каково значение меры качества *Accurасу* для такой модели?

Ответ: для такой (простой и интуитивно глупой) модели доля верных ответов:

$$Acc = 0,95 = 95\%$$

Хорошая ли это мера качества для такой (несбалансированной) выборки?

Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

Самая простая и (чаще всего) неверная мера качества в задачах классификации – доля верных ответов (accuracy)

$$Accuracy(y_{pred}, y_{true}) = \frac{\sum[y_{pred} == y_{true}]}{N}$$

Представим следующую задачу (и решение):

В наборе данных по диагностике онкологических заболеваний \mathcal{T} 99.5% объектов – класса «здоров», остальные объекты – классов B,C,D (различные виды злокачественных новообразований)

Мы создали и обучили модель, которая обладает очень высокой чувствительностью: не пропускает ни одного случая злокачественного новообразования. Но при этом доля ложноположительных диагнозов довольно высока: 2% диагностируемых пациентов.

Каково значение меры качества *Accurасу* для такой модели?

Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

Представим следующую задачу (и решение):

В наборе данных по диагностике онкологических заболеваний \mathcal{T} 99.5% объектов – класса «здоров», остальные объекты – классов B,C,D (различные виды злокачественных новообразований)

Ответ нашей модели		
	НЕТ	ДА
НЕТ		
ДА		

Мы создали и обучили модель, которая обладает очень высокой чувствительностью: не пропускает ни одного случая злокачественного новообразования. Но при этом доля ложноположительных диагнозов довольно высока: 2% диагностируемых пациентов.

Каково значение меры качества *Accurасу* для такой модели?

Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

Представим следующую задачу (и решение):

Ответ нашей модели				1000
		НЕТ	ДА	
ground truth	НЕТ	975	20	995
	ДА	0	5	5

В наборе данных по диагностике онкологических заболеваний \mathcal{T} 99.5% объектов – класса «здоров», остальные объекты – классов B,C,D (различные виды злокачественных новообразований)

Мы создали и обучили модель, которая обладает очень высокой чувствительностью: не пропускает ни одного случая злокачественного новообразования. Но при этом доля ложноположительных диагнозов довольно высока: 2% диагностируемых пациентов.

Каково значение меры качества *Accurасу* для такой модели?

Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

Представим следующую задачу (и решение):

В наборе данных по диагностике онкологических заболеваний \mathcal{T} 99.5% объектов – класса «здоров», остальные объекты – классов B,C,D (различные виды злокачественных новообразований)

		Ответ нашей модели		
		НЕТ	ДА	1000
ground truth	НЕТ	975	20	995
	ДА	0	5	5

Мы создали и обучили модель, которая обладает очень высокой чувствительностью: не пропускает ни одного случая злокачественного новообразования. Но при этом доля ложноположительных диагнозов довольно высока: 2% диагностируемых пациентов.

Каково значение меры качества *Accurасy* для такой модели?

$$Acc = \frac{975 + 5}{1000} = 0.98 = 98\%$$

Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

Представим следующую задачу (и решение):

Ответ нашей модели			
		НЕТ	ДА
ground truth	НЕТ	TN	FP
	ДА	FN	TP

1000 995 5

В наборе данных по диагностике онкологических заболеваний \mathcal{T} 99.5% объектов – класса «здоров», остальные объекты – классов B,C,D (различные виды злокачественных новообразований)

Мы создали и обучили модель, которая обладает очень высокой чувствительностью: не пропускает ни одного случая злокачественного новообразования. Но при этом доля ложноположительных диагнозов довольно высока: 2% диагностируемых пациентов.

Каково значение меры качества *Accurasy* для такой модели?

$$Acc = 0,98$$

TP – True Positive, доля верно определенных ответов класса «1» («Истина»)

TN – True Negative, доля верно определенных ответов класса «0» («Ложь»)

FP – False Positive (False alarms), доля ответов класса 0, ошибочно классифицированных как положительные

FN – False Negative (Misses), доля ответов класса 1, ошибочно классифицированных как отрицательные

Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

		Ответ нашей модели		Negative Predictive Value
		НЕТ	ДА	1000
ground truth	НЕТ	TN 975	FP 20	995
	ДА	FN 0	TP 5	5

$$NPV = \frac{TN}{TN + FN}$$

$$NPV = \frac{975}{975 + 0} = 1,0$$

Какая доля объектов, идентифицированных (тестом/алгоритмом/моделью) как объекты класса «0», действительно имеют класс «0».

С какой вероятностью пациент действительно здоров, если тест выдал отрицательный результат.

Когда **NPV близок к нулю** (если **доля FN велика**), наша модель на предоставленных данных «предпочитает» выдавать ложноотрицательный ответ вместо положительного – то есть, **«по умолчанию пациент скорее здоров»**. Относительно положительного результата теста это консервативная оценка.

Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

		Ответ нашей модели		Positive Predictive Value, PPV, Precision, Точность
		НЕТ	ДА	1000
ground truth	НЕТ	TN 975	FP 20	995
	ДА	FN 0	TP 5	5

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

$$P = \frac{5}{5 + 20} = 0,2$$

Какая доля объектов, идентифицированных как класс «1», действительно имеют класс «1»

С какой вероятностью пациент действительно болен, если тест выдал положительный результат.

Когда **точность близка к нулю (доля FP велика)**, наша модель на предоставленных данных «предпочитает» выдавать ложноположительный ответ вместо отрицательного – то есть, «по умолчанию пациент скорее болен».

Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

		Ответ нашей модели		Чувствительность, полнота, Sensitivity, Recall, True Positive Rate, TPR
		НЕТ	ДА	1000
ground truth	НЕТ	TN 975	FP 20	995
	ДА	FN 0	TP 5	5

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

$$R = \frac{5}{0 + 5} = 1,0$$

Какая доля объектов класса «1» определяются (тестом/алгоритмом/моделью) как имеющие класс «1»?

С какой вероятностью тест даст положительный результат, если пациент болен?

Когда чувствительность (теста, модели, алгоритма) близка к нулю (доля FN велика), наша модель на предоставленных данных пропускает слишком много «положительных» объектов – тест/модель/алгоритм не слишком чувствителен

Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

		Ответ нашей модели		Специфичность, Specificity, True Negative Rate, TNR
		НЕТ	ДА	1000
ground truth	НЕТ	TN 975	FP 20	995
	ДА	FN 0	TP 5	5

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

$$\text{Specificity} = \frac{975}{975 + 20} = 0,98$$

Какая доля объектов класса «0» определяются (тестом/алгоритмом/моделью) как имеющие класс «0»?

С какой вероятностью тест даст отрицательный результат, если пациент здоров?

Когда специфичность теста/модели/алгоритма близка к нулю (доля FP велика), наша модель на предоставленных данных слишком часто выдает положительный ответ(диагноз) в тех случаях, когда объект на самом деле класса «0». Такой тест (такая модель, такой алгоритм) не слишком специфичен для решаемой задачи.

Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

F_β -мера

Ответ нашей модели

		НЕТ	ДА	1000
ground truth	НЕТ	TN 975	FP 20	995
	ДА	FN 0	TP 5	5

$$F_\beta = (1 + \beta^2) \frac{P * R}{(\beta^2 * P) + R}$$

$$F_1 = 2 \frac{P * R}{P + R}$$

$$F_1 = 2 * \frac{0.2 * 1}{0.2 + 1} = 0,333$$

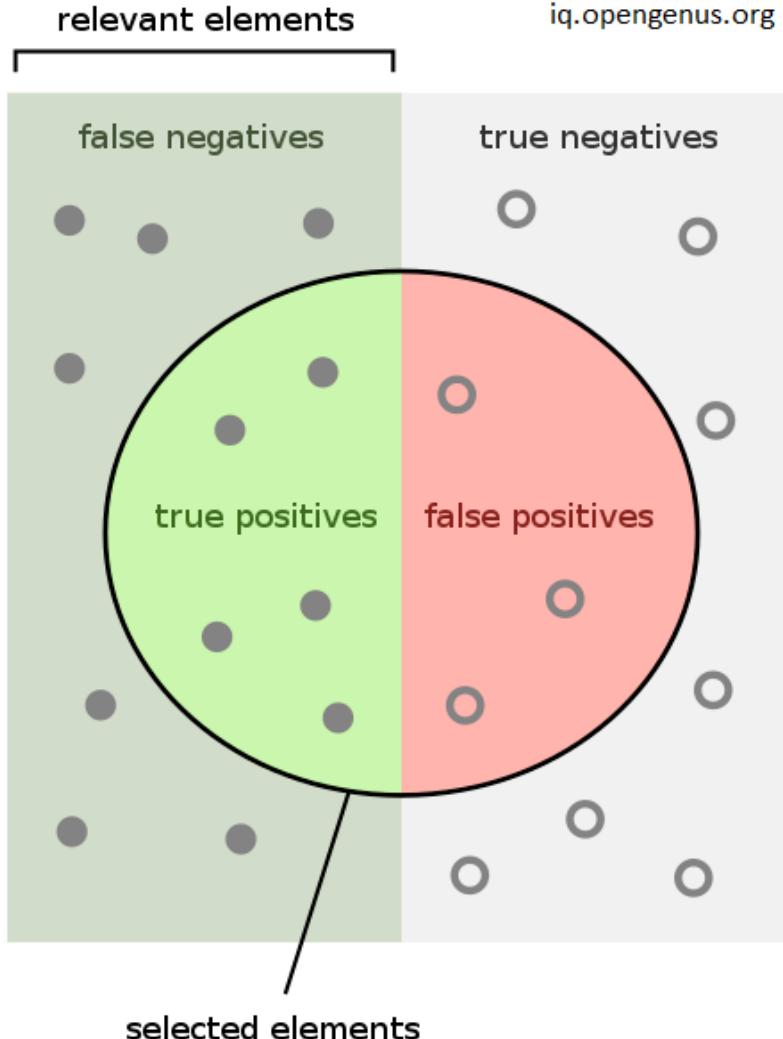
F_1 -мера – среднее гармоническое точности и полноты

Точность: $P = \frac{TP}{TP+FP}$

Полнота: $R = \frac{TP}{TP+FN}$

Оценка качества моделей классификации

Меры качества в задаче бинарной классификации



How many relevant items are selected?
e.g. How many sick people are correctly identified as having the condition.

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

(Recall)



How many negative selected elements are truly negative?
e.g. How many healthy people are identified as not having the condition.

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

(Precision)



Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

AUC ROC

FPR

		Ответ нашей модели		1000
		НЕТ	ДА	
ground truth	НЕТ	TN 975	FP 20	995
	ДА	FN 0	TP 5	5

TPR (Recall)

		Ответ нашей модели		1000
		НЕТ	ДА	
ground truth	НЕТ	TN 975	FP 20	995
	ДА	FN 0	TP 5	5

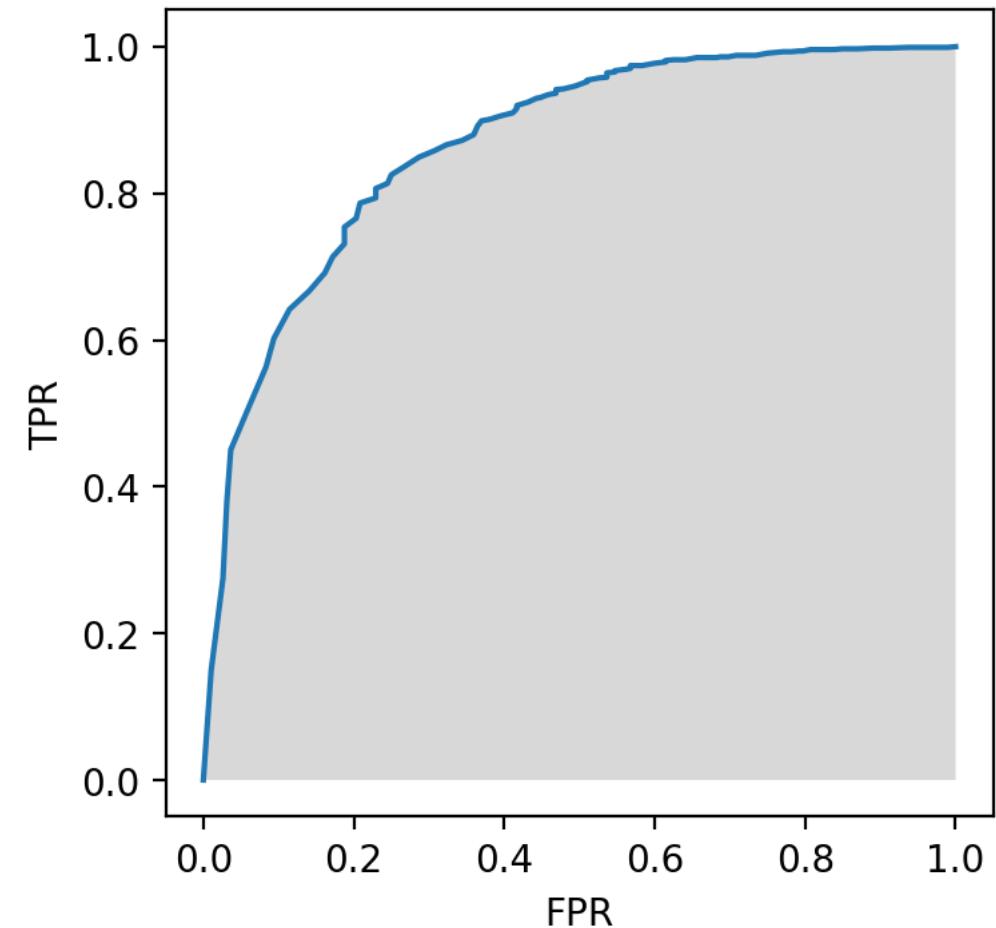
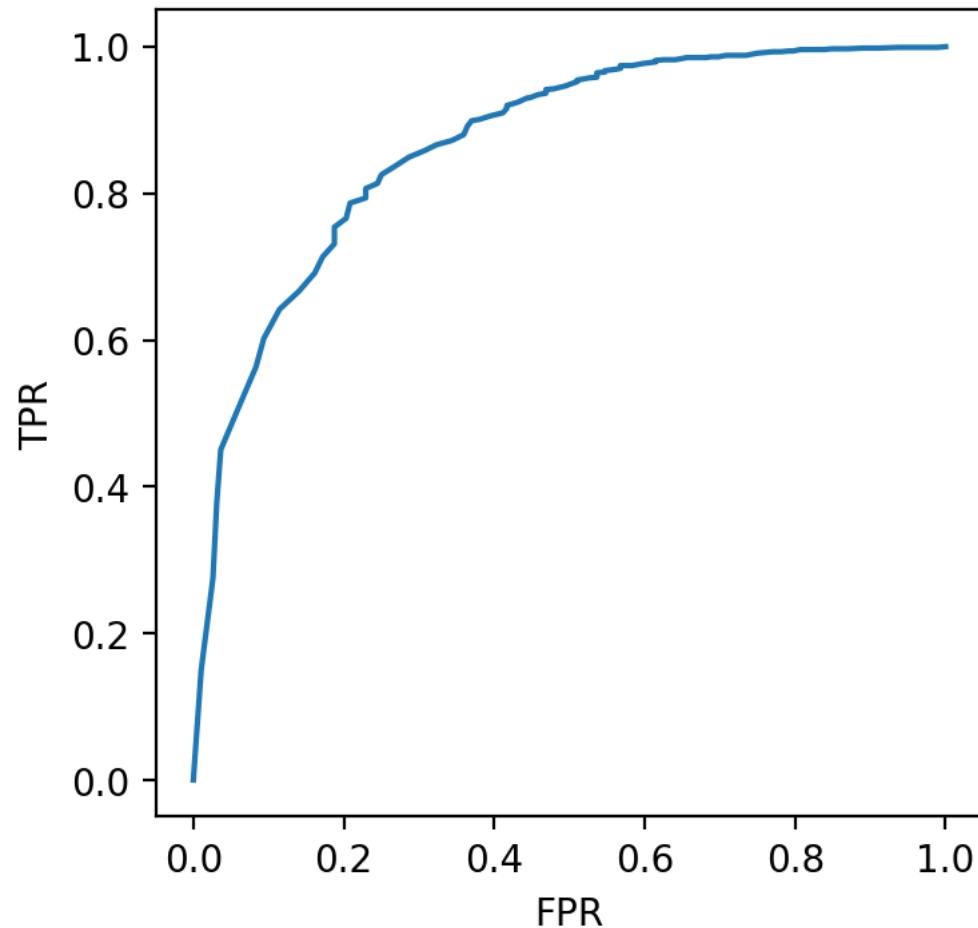
$$FPR = \frac{FP}{FP + TN}$$

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

Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

AUC ROC



Оценка качества моделей классификации

Меры качества в задаче бинарной классификации

AUC ROC

- AUC ROC характеризует метод/алгоритм/тест в целом, а не конкретную реализацию с заданным пороговым значением вероятности;
- AUC ROC инвариантен к пороговому значению вероятности;
- AUC ROC инвариантен к масштабу оценки вероятности (зависит от ранжирования объектов, но не абсолютных значений вероятностей);

Mindray IgG-антитела к антигенам вируса SARS-CoV-2 выявлены в 165 образцах (99,4%), на «Вектор-Бест» – в 164 сыворотках (98,8%), на «Диагностических системах» – в 151 (90,96%), на Хема – в 154 (92,8%), а на Abbott – в 155 образцах (93,4%). При этом 135 (81,33%) образцов были положительными во всех тест-системах, тогда как 30 образцов имели дискордантные результаты (18,07%), а в 9 сыворотках специфических IgG не обнаруживалось в 2 и более тест-системах. ROC-анализ выявил высокую диагностическую ценность всех исследованных тест-систем (AUC от 0,908 до 0,998), что свидетельствует о высоком качестве разделительной модели положительных и отрицательных образцов ($p < 0,001$). При заданных производителями тест-систем cut-off чувствительность и специфичность находилась в диапазоне от 82,8% и 93,3% для набора