



Машинное обучение в науках о Земле

Михаил Криницкий

к.т.н., н.с. Институт океанологии РАН им. П.П. Ширшова

Лаборатория взаимодействия океана и атмосферы и мониторинга климатических изменений (ЛВОАМКИ)





Задачи классификации Оценка качества моделей классификации

Михаил Криницкий

K.T.H., H.C.

Институт океанологии РАН им. П.П. Ширшова

Лаборатория взаимодействия океана и атмосферы и мониторинга климатических изменений (ЛВОАМКИ)

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 A, остальные объекты — классов B,C,D Мы создали и обучили модель, которая для любого нового объекта выдает результат «это объект класса A»

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 A, остальные объекты — классов B,C,D Мы создали и обучили модель, которая для любого нового объекта выдает результат «это объект класса A»

Каково значение меры качества *Accuracy* для такой модели?

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

$$Acc = 95\%$$

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.

Меры качества в задаче бинарной классификации

Ответ нашей модели

HET ДА

HET ДА

ДА

Представим следующую задачу (и решение):

В наборе данных по диагностике онкологических заболеваний \mathcal{T} 99.5% объектов — класса «здоров», остальные объекты — классов В,С,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.

Меры качества в задаче бинарной классификации

Ответ нашей модели

		HET	ДА	1000
ground truth	HET	975	20	995
	ДА	0	5	5

Представим следующую задачу (и решение):

В наборе данных по диагностике онкологических заболеваний \mathcal{T} 99.5% объектов — класса «здоров», остальные объекты — классов В,С,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.

Меры качества в задаче бинарной классификации

Ответ нашей модели

HET ДА 1000

HET 975 20 995

ДА 0 5 5

ground

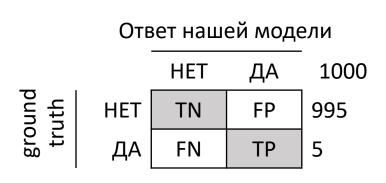
Представим следующую задачу (и решение):

В наборе данных по диагностике онкологических заболеваний \mathcal{T} 99.5% объектов — класса «здоров», остальные объекты — классов В,С,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.

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

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 В,С,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 *Accuracy* для такой модели?

$$Acc = 98\%$$

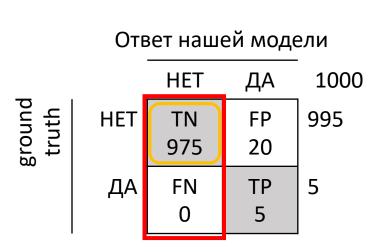
TP – True Positive, доля верно определенных ответов класса «1» («Истина»)

TN – True Negative, доля верно определенных ответов класса «О» («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

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 В,С,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 *Accuracy* для такой модели?

$$Acc = 98\%$$

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

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

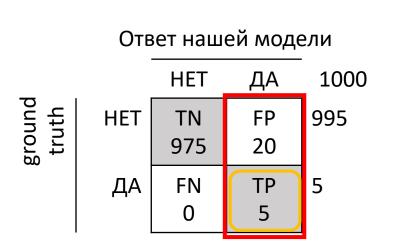
FP – False Positive (False alarms), доля ответов класса 0, ошибочно классифицированных как положительные

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

• Negative Predictive Value,
$$NPV = \frac{TN}{TN+FN} = \frac{975}{975+0} = 100\%$$

Когда NPV близок к нулю (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.

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 В,С,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 *Accuracy* для такой модели?

$$Acc = 98\%$$

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

• Positive Predictive Value, PPV, Precision (точность), $P = \frac{TP}{TP + FP} = \frac{5}{20 + 5} = 20\%$.

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

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 В,С,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 *Accuracy* для такой модели?

$$Acc = 98\%$$

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

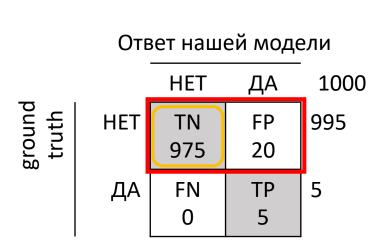
FN – False Negative (Misses), доля ответов класса 1, ошибочно классифицированных как отрицательные

• Чувствительность, полнота, Recall, True Positive Rate, TPR,

$$R = \frac{TP}{TP + FN} = \frac{5}{0+5} = 100\%$$

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

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 В,С,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 *Accuracy* для такой модели?

$$Acc = 98\%$$

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

• Специфичность, True Negative Rate, TNR, Specificity $=\frac{TN}{TN+FP}=\frac{975}{975+20}=98\%$

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.

Меры качества в задаче бинарной классификации

Представим следующую задачу (и решение):

Ответ нашей модели

HET ДА 1000

HET TN FP 995

975 20

ДА FN TP 5
0 5

В наборе данных по диагностике онкологических заболеваний \mathcal{T} 99.5% объектов — класса «здоров», остальные объекты — классов В,С,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 *Accuracy* для такой модели?

$$Acc = 98\%$$

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

• F-mepa:
$$F_{\beta} = (1 + \beta^2) \frac{P*R}{(\beta^2*P)+R} = \{\beta = 1\} = \frac{P*R}{P+R}; \qquad F_1 = \frac{0.2*1}{0.2+1} = 0.167$$