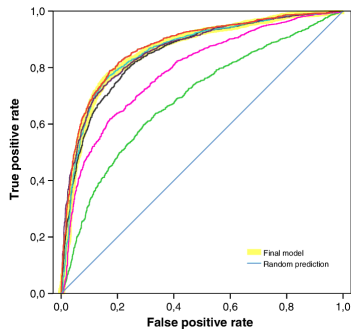


Оценка классификаторов

Виктор Китов

v.v.kitov@yandex.ru



Содержание

- 1 Оценка прогнозов меток классов
- 2 Оценка прогнозов меток и вероятностей
- 3 ROC кривые

Матрица ошибок

Матрица ошибок (confusion matrix) - таблица сопряженности между y и \hat{y} :

		Прогноз \hat{y}			
		1	2	...	C
Истинный класс y	1	n_{11}	n_{12}		
	2	n_{21}	n_{22}		
	\vdots			\ddots	
	C				n_{CC}

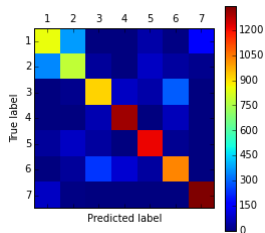
Корректные классификации - на диагонали.

Ошибки классификации - вне диагонали.

$$\text{Accuracy} = \frac{\sum_{c=1}^C n_{cc}}{\sum_{i,j=1}^C n_{ij}}; \quad \text{ErrorRate} = 1 - \text{Accuracy} = \frac{\sum_{i,j=1; i \neq j}^C n_{ij}}{\sum_{i,j=1}^C n_{ij}}$$

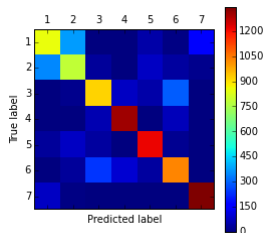
Визуализация матрицы ошибок

Визуализация матрицы ошибок (географическая цветовая схема)



Визуализация матрицы ошибок

Визуализация матрицы ошибок (географическая цветовая схема)



- Видим, что ошибки сконцентрированы на разделении классов 1 и 2.
- Вариант решения:
 - объединим классы 1 и 2 в новый класс «1+2»
 - решим задачу классификации на $\{\text{«1+2»}, 3, 4, 5, 6, 7\}$
 - разделим множество «1+2» отдельным классификатором.

Случай 2х классов

Матрица ошибок:

		Прогноз		
		+	-	всего
Факт	+	TP (true positives)	FN (false negatives)	P
	-	FP (false positives)	TN (true negatives)	N
	всего	\hat{P}	\hat{N}	

Случай 2х классов

Матрица ошибок:

		Прогноз		всего
		+	-	
Факт	+	TP (true positives)	FN (false negatives)	P
	-	FP (false positives)	TN (true negatives)	N
всего		\hat{P}	\hat{N}	

Точность:	$\frac{TP+TN}{P+N}$
Частота ошибок:	1-точность = $\frac{FP+FN}{P+N}$

Случай 2х классов

Матрица ошибок:

		Прогноз		всего
		+	-	
Факт	+	TP (true positives)	FN (false negatives)	P
	-	FP (false positives)	TN (true negatives)	N
всего		\hat{P}	\hat{N}	

Точность:	$\frac{TP+TN}{P+N}$
Частота ошибок:	$1\text{-точность} = \frac{FP+FN}{P+N}$

Точность и частота ошибок не информативны для неравномерного распределения классов.

Метрики качества для положительного класса

Точность (Precision)	$\frac{TP}{\hat{P}}$
Полнота (Recall), TPR	$\frac{TP}{P}$
F-мера	$\frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$
Взвешенная F-мера	$\frac{1}{\frac{\beta^2}{1+\beta^2} \frac{1}{Precision} + \frac{1}{1+\beta^2} \frac{1}{Recall}}$

TPR, recall	$\frac{TP}{P}$
FPR	$\frac{FP}{N}$

- Доля правильных положительных классификаций TPR
 - true positive rate, recognition rate
- Доля неправильных положительных классификаций FRP
 - false positive rate, false alarm.

Содержание

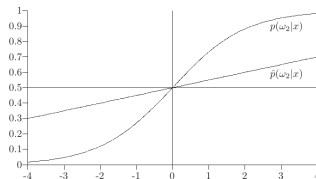
- 1 Оценка прогнозов меток классов
- 2 Оценка прогнозов меток и вероятностей
- 3 ROC кривые

Оценка прогнозов меток и вероятностей

- **Дискриминационные (discriminability) метрики качества** оценивают качество предсказания меток классов.
 - примеры: частота ошибок, точность, полнота, и т.д.

Оценка прогнозов меток и вероятностей

- **Дискриминационные (discriminability) метрики качества** оценивают качество предсказания меток классов.
 - примеры: частота ошибок, точность, полнота, и т.д.
- **Вероятностные (reliability) метрики качества** оценивают качество предсказания вероятностей классов.
 - условное правдоподобие выборки $\prod_{i=1}^N \hat{p}(y_i|x_i)$
 - оценка Бриера (Brier score): $\frac{1}{N} \sum_{n=1}^N \|\mathbf{p}_n - \hat{\mathbf{p}}_n\|^2$



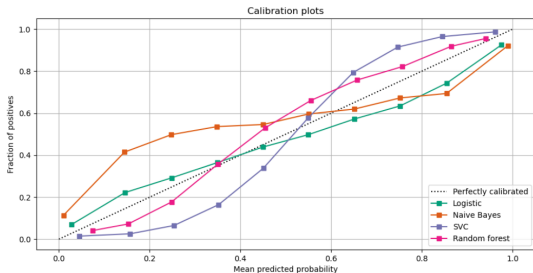
Пример: метки прогнозируются хорошо, а классы плохо.

Предсказание вероятностей

- Любой классификатор можно научить предсказывать вероятности классов подобрав $f(\cdot)$:

$$p(y = +1|x) = f(g(x))$$

- График калибровки (calibration plot) - соответствие между $p(y = +1|x)$ и $f(g(x_n))$ для объектов из ячеек $\{\beta_i \leq f(g(x)) \leq \beta_{i+1}\}_i$:



Основные подходы для предсказания вероятностей

- шкалирование Платта
 - A, B находятся методом максимального правдоподобия по отложенной выборке

$$p(y = +1|x) = \frac{1}{1 + \exp(Ag(x) + B)}$$

- напрямую по графику калибровки (гистограмме)
- с помощью изотонической регрессии (isotonic regression)

$$\begin{cases} \sum_{n=1}^N (\hat{y}_n - y_n)^2 \rightarrow \min_{\hat{y}_1, \dots, \hat{y}_N} \\ \hat{y}_j \geq \hat{y}_i \quad \forall (i, j) : x_j \geq x_i \end{cases}$$

Содержание

- 1 Оценка прогнозов меток классов
- 2 Оценка прогнозов меток и вероятностей
- 3 ROC кривые**

Решающее правило бинарной классификации

- Используем относительный рейтинг
 $g(x) = g_{+1}(x) - g_{-1}(x)$.
- Классификация $\hat{y}(x) = \text{sign}(g(x))$.
- Введем параметр $\alpha \in \mathbb{R}$, контролирующий предпочтения между классами:

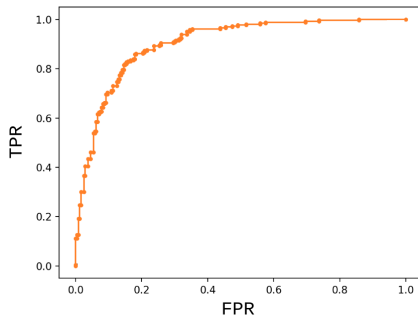
$$\hat{y}(x) = \text{sign}(g(x) - \alpha)$$

- $\downarrow \alpha$: больше $\hat{y} = +1$; $\uparrow \alpha$: меньше $\hat{y} = +1$.
- Можем обучить модель 1 раз, а потом использовать в разных режимах:
 - детекция самолетов в мирное/военное время
 - выдача кредитов в период экономического бума/спада
- В случае неравных потерь $\lambda_{+1} \neq \lambda_{-1}$ тоже нужно подбирать α :
 - $\lambda_{+1} = \text{cost}(\hat{y} = -1 | y = +1)$; $\lambda_{-1} = \text{cost}(\hat{y} = +1 | y = -1)$

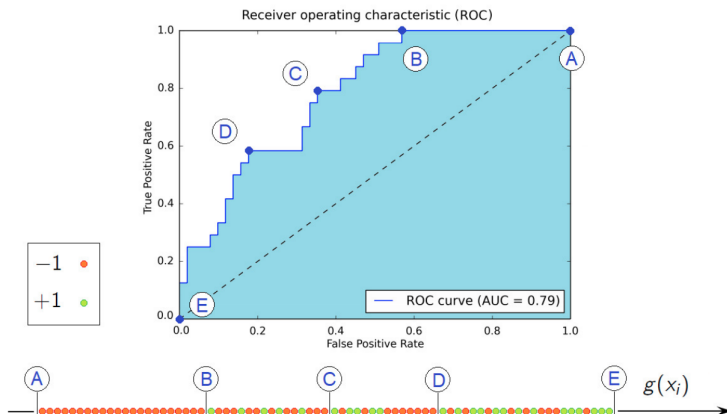
ROC кривая¹

- $TPR = TPR(\alpha)$, $FPR = FPR(\alpha)$.
- ROC кривая- функция $TPR(FPR)$.
 - receiver operating characteristic (рабочая характеристика приёмника)

$$TPR = \frac{TP}{P} \quad FPR = \frac{FP}{N}$$



Построение ROC-кривой по выборке



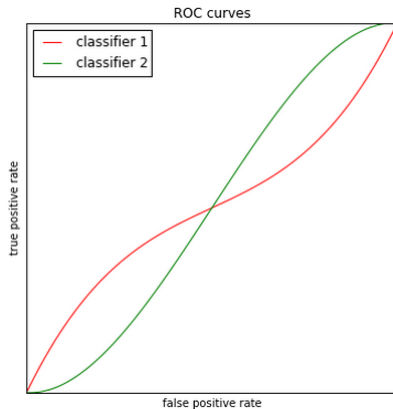
Вопросы

- Более высота ROC кривой связана с качеством классификации?
- Какова ROC-кривая для случайного угадывания $\hat{y}(x) = \text{sign}(\xi - \alpha)$, $\xi \sim \text{Uniform}[0, 1]$?
- Как улучшить классификатор для вогнутой ROC кривой?
- Как поменяется ROC кривая при инвертировании классификатора:

$$\text{sign}(g(x) - \alpha) \longrightarrow \text{sign}(\alpha - g(x))$$

Композиции классификаторов

Как создать семейство классификаторов с максимально высокой ROC-кривой в таком случае?

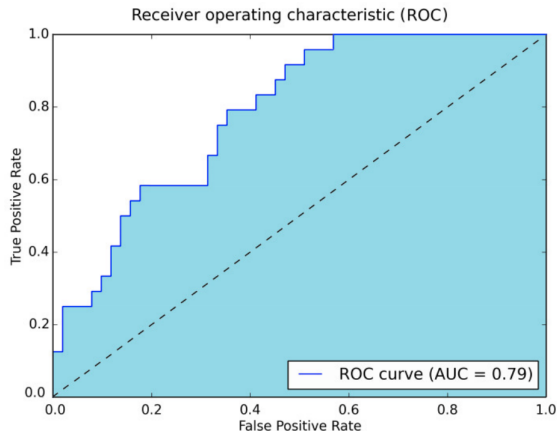


Преимущества ROC-кривой

- оценка семейства классификаторов, параметризованных α .
- инвариантность к монотонным преобразованиям
 $g(x) \rightarrow g'(x) = f(g(x)), \uparrow f(\cdot)$
 - $\text{sign}(g(x) > \alpha) \iff \text{sign}(f(g(x)) > f(\alpha))$, поэтому точке $TPR(\alpha), FPR(\alpha)$ соответствует $TPR'(f(\alpha)), FPR'(f(\alpha))$.

Площадь под кривой

- Площадь под ROC-кривой (area under curve, AUC) - интегральная мера качества семейства классификаторов.



Эквивалентное определение AUC

- **AUC=доле корректно упорядоченных пар:**
- выберем случайно
 - отрицательный объект ($x_i, y_i = -1$) и положительный объект ($x_j, y_j = +1$)
 - тогда AUC - вероятность верного упорядочивания таких объектов

$$AUC = p(g(x_i) < g(x_j))$$

AUC=доля верно упорядоченных пар объектов

$x_{(1)}, \dots, x_{(N)}$ - упорядоченные объекты по рейтингу

$$g(x_{(1)}) < g(x_{(2)}) < \dots < g(x_{(N)}), \quad \hat{y}_k(x) = \text{sign}(g(x) \geq g(x_{(k)}))$$

$$TPR_k = \frac{\sum_{n=k}^N \mathbb{I}[y_{(n)} = +1]}{N_+}, \quad FPR_k = \frac{\sum_{n=k}^N \mathbb{I}[y_{(n)} = -1]}{N_-}$$

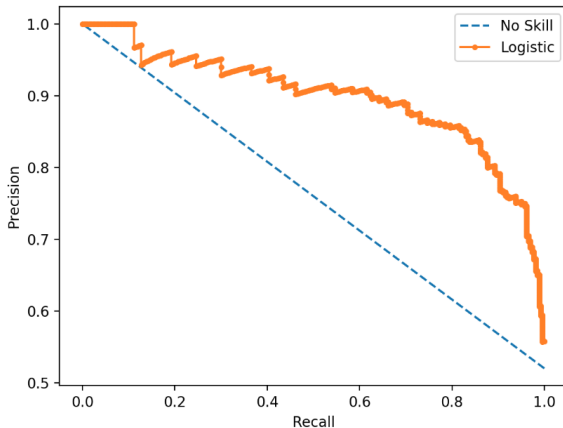
$$AUC = \sum_{k=1}^{N-1} \frac{TPR_{k+1} + TPR_k}{2} (FPR_{k+1} - FPR_k) =$$

$$= \sum_{k=1}^{N-1} \frac{\sum_{n=k+1}^N \mathbb{I}[y_{(n)} = +1] + \frac{1}{2} \mathbb{I}[y_{(k)} = +1]}{N_+} \cdot \frac{\mathbb{I}[y_{(k)} = -1]}{N_-} =$$

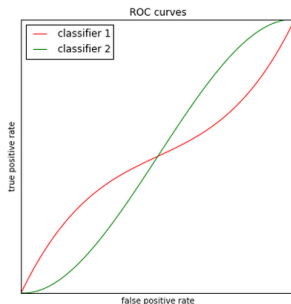
$$= \frac{1}{N_+ N_-} \sum_{k=1}^{N-1} \sum_{n=k+1}^N \mathbb{I}[y_{(n)} = +1] \mathbb{I}[y_{(k)} = -1] = \frac{1}{N_+ N_-} \sum_{k < n} \mathbb{I}[y_{(k)} < y_{(n)}]$$

Аналог ROC-кривой - precision(recall)

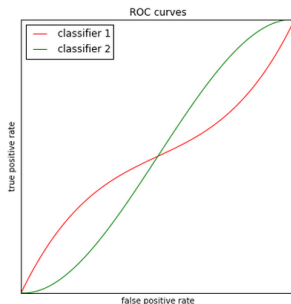
$$Precision = \frac{TP}{\widehat{P}} \quad Recall = \frac{TP}{P}$$



Сравнение классификаторов по ROC-кривой



Сравнение классификаторов по ROC-кривой



Как сравнивать классификаторы?

- Фиксированные $\lambda_{+1}, \lambda_{-1}$: по точкам на ROC-кривой
- Неизвестные $\lambda_{+1}, \lambda_{-1}$: по AUC
- Частично известные $\lambda_{+1}, \lambda_{-1}$: по LC-индексу

Изолинии потерь

- Вероятности ошибок: $p(\hat{y} = -1|y = +1) = 1 - TPR$,
 $p(\hat{y} = +1|y = -1) = FPR$
- Изолиния потерь (потери $\equiv L$):

$$\begin{aligned}L &= p(y = +1)(1 - TPR)\lambda_{+1} + p(y = -1)FPR\lambda_{-1} \\(TPR - 1)p(y = +1)\lambda_{+1} &= -L + \lambda_{-1}p(y = -1)FPR \\TPR &= 1 + \frac{\lambda_{-1}p(y = -1)FPR - L}{\lambda_{+1}p(y = +1)}\end{aligned}$$

- Оптимальная точка: в точке касания изолинии к ROC-кривой
 - это точка на ROC-кривой с тангенсом угла наклона

$$tg = \frac{\lambda_{-1}p(y = -1)}{\lambda_{+1}p(y = +1)}$$

Площадь под кривой (AUC)

- Глобальная характеристика качества для различных α
- $AUC \in [0, 1]$
- $AUC=0.5$ - случайное угадывание
- $AUC=1$ - идеальное упорядочивание (безошибочная классификация при определенном α)

LC индекс

- Заданные $\lambda_{+1}, \lambda_{-1}$: слишком специфично.
 - Неопределенные $\lambda_{+1}, \lambda_{-1}$: слишком общий случай.
 - LC индекс вычисляют для промежуточного сценария.
- 1 отмасштабируем λ_{+1} и λ_{-1} так, что $\lambda_{+1} + \lambda_{-1} = 1$
 - 2 определим $\lambda_1 = \lambda, \lambda_{-1} = 1 - \lambda$
 - 3 для каждого $\lambda \in [0, 1]$ вычислим
$$S(\lambda) = \begin{cases} +1 & \text{если 1й классификатор лучше} \\ -1 & \text{если 2й классификатор лучше} \end{cases}$$
 - 4 прикинем плотность распределения λ : $p(\lambda)$ (например, "треугольный" случай)
 - 5 выберем классификатор 1, если $\int_0^1 S(\lambda)p(\lambda)d\lambda > 0$ иначе классификатор 2.

Точность и полнота - многоклассовый случай

	бинарный случай	макроусреднение	микроусреднение
Точность	$\frac{TP}{\hat{P}}$	$\frac{1}{C} \sum_{c=1}^C \frac{TP_c}{\hat{P}_c}$	$\frac{\sum_{c=1}^C TP_c}{\sum_{c=1}^C \hat{P}_c}$
Полнота	$\frac{TP}{P}$	$\frac{1}{C} \sum_{c=1}^C \frac{TP_c}{P_c}$	$\frac{\sum_{c=1}^C TP_c}{\sum_{c=1}^C P_c}$

Обозначения:

- TP_c - # верно предсказанных объектов класса c .
- P_c - # объектов класса c .
- \hat{P}_c - # объектов, предсказанных как класс c .

Макроусреднение учитывает метрики равномерно по классам (чувствительно к классам с малым числом объектов), а микроусреднение - равномерно по объектам.

Заключение

- Матрица ошибок дает больше информации, чем частота ошибок.
- Precision, recall и (TPR, FPR) используются для несбалансированных классов.
 - многоклассовый случай: микро или макроусреднение
- Можно оценивать качество
 - предсказания меток классов (accuracy, precision, recall)
 - упорядочивания по рейтингу (ROC-кривая, AUC)
 - вероятностей классов (Оценка Бриера, условное правдоподобие)
- Площадь под ROC-кривой-вероятность верного упорядочивания всех пар объектов по рейтингу.
- LC индекс лучше AUC, когда есть частичная информация о $\lambda_{+1}, \lambda_{-1}$.