Машинное обучение в финансах

Лекция 10: Логистическая регрессия

Роман В. Литвинов*

*CRO
Финансовая Группа БКС

Высшая школа экономики, Май 2021

Содержание

Содержание

Классификация

Вторая основная задача машинного обучения, как мы говорили на шестой лекции, это задача классификации.

Мы помним, что задачи классификации сводятся к поиску функции $f : \mathbb{R}^n \to [0, 1..., m]$. Т.е. к задаче присваивания определенной категории/ класса объекту, где m - этот общее количество таких классов. В простейшем случае бинарной классификации (binary classification) - таких категорий две (0 или 1).

Интуитивно понятная аналогия - разбиение объектов по m корзинам, наклеивание m ярылыков на объекты и тп.

Как мы с вами увидим, многие методы классификации, базируются на основе оценки вероятности принадлежности объекта к тому или иному классу (непрерывная величина на отрезке от 0 до 1) и в этом смысле чем-то похожи на регрессионные модели. Сначала присваивается вероятность, а затем решается на основе этой оценки вероятности, принадлежит объект к этому или же другому классу.

Содержание

Простая логистическая регрессия

Простая логистическая регрессия подразумевает бинарную классификацию (1 или 0, дефолт или не дефолт, белое или не белое).

Понятно, что "лобовая"попытка использования линейной регрессии для оценки значения классификатора работать корректно не будет.

Основной фокус здесь, это не оценка значения классификатора, а оценка вероятности того, что наблюдение принадлежит соответствующему классу.

Пусть $y \in$ классу 1 с вероятностью p и, соответственно, $y \notin$ классу 1 с вероятностью 1-p (т.е. принадлежит классу 0).

Наша задача будет сводится к оценке вероятности p - корректному прогнозу значения этой вероятности. Имея на руках такой прогноз, мы можем условиться, что все наблюдения имеющие вероятность $p > 0.5$ будут отнесены к 1-му классу. Такое значение p называется порогом решающего правила (treshold).

Простая логистическая регрессия

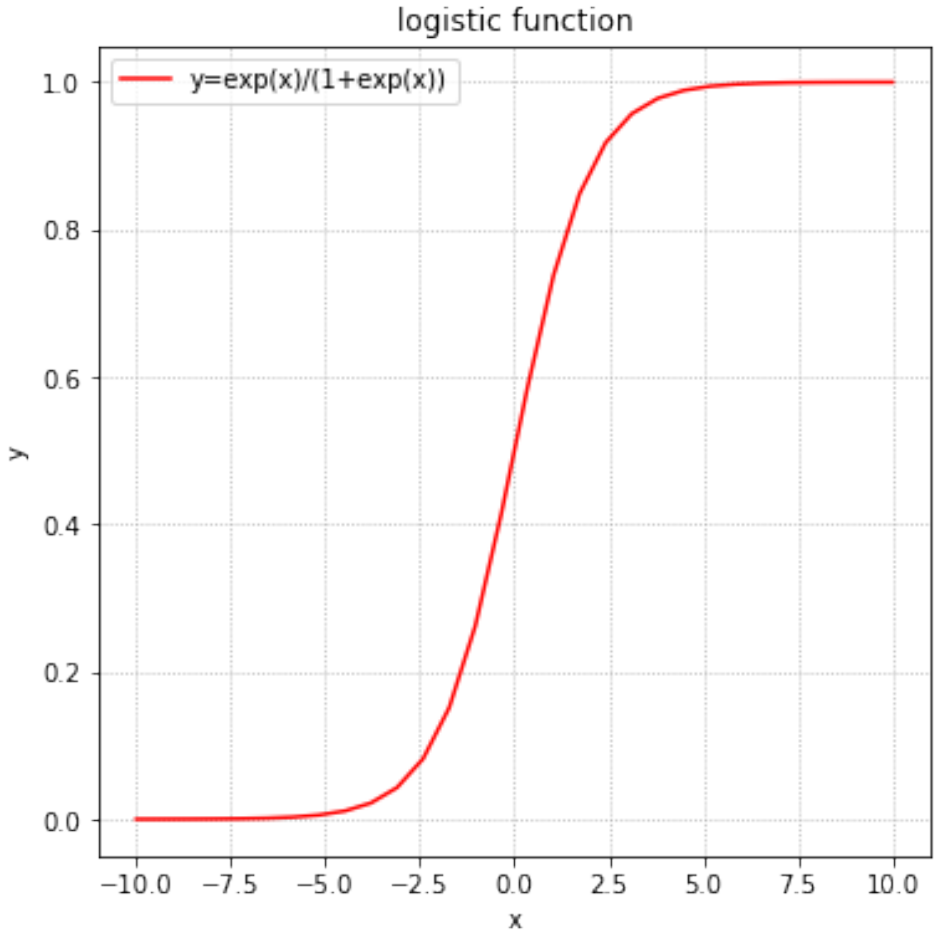
Прежде чем мы двинемся к схеме оценки этой вероятности, необходимо отметить тот факт, что, как мы знаем, вероятность может принимать значения от 0 до 1. Т.о. простое применение методов линейной регрессии к оценке нашего параметра p снова не будет работать корректно.

Нам понадобится понятие логистической функции. Стандартная логистическая функция (standard logistic function) представляет собой функцию вида:

$$f(x) = \frac{e^x}{1 + e^x}$$

При области определения x от $-\infty$ до $+\infty$ функция будет иметь обл. значений от 0 до 1. Значение $f(x)$ будет стремиться к нулю по мере приближения x к $-\infty$, и к единице - по мере приближения x к $+\infty$.

Простая логистическая регрессия



Простая логистическая регрессия

В простой логистической регрессии мы будем использовать следующую модификацию логистической функции. Вероятность принадлежности наблюдения к заданному классу будет следующим образом зависеть от фактора/ предиктора:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{e^{\beta_0+\beta_1 Factor_1}}{1+e^{\beta_0+\beta_1 Factor_1}}$$

Наша задача - подобрать коэффициенты β_0 и β_1 таким образом, чтобы предсказываемая для i-того наблюдения вероятность принадлежности к заданному классу была максимально близка к наблюдаемому в реальности классу.

Что это означает на практике?

Простая логистическая регрессия

У нас есть i-наблюдений (наши сигналы/доходности, напр. + или -). Обозначим их как r_i. Мы имеем два класса, к которому в реальности относится каждое наблюдение - 0 и 1. Т.е. $r_i = 1$ или $r_i = 0$.

Для каждого наблюдения мы вычисляем предсказанную величину вероятности того, что наблюдение относится к:

- классу 1, т.е. $p(r_i = 1)$
- классу 0 (т.е. не отн. к классу 1), т.е. $p(r_i = 0) = 1 - p(r_i = 1)$.

В идеальном мире для i-того наблюдения r_i, относящегося, например, к классу 1 ($r_i = 1$) наша модель должна давать оценку вероятности того, что данное наблюдение относится к заданному классу равную единице: $p(r_i = 1) = 1$.

В реальности, наша задача сводится к тому, чтобы подобрать коэффициенты β_0 и β_1 т.о., чтобы эта вероятность была максимально близкой к единице (к нулю для наблюдений относящихся к классу 0) - минимизировать ошибку между фактом и прогнозом.

Содержание

Метод максимального правдоподобия

Для поиска оптимальных весов нашей модели (β_0, β_1) обычно используется метод максимального правдоподобия (maximum likelihood estimation или MLE).

Сутью этого метода является максимизация функции правдоподобия (likelihood function). Эта функция характеризует вероятность наблюдения имеющегося набора данных (выборки) в зависимости от определенных параметров. Максимальное значение функции будет говорить о том, что именно этот набор данных, которые мы имеем на руках, наиболее вероятен при заданных параметрах функции (коэффициентах β).

Наша задача - подобрать коэффициенты β_0 и β_1 таким образом, чтобы функция правдоподобия имела максимальное значение (т.е. при таких параметрах именно наблюдаемая выборка была наиболее вероятной).

Метод максимального правдоподобия

Мы начнем построение функции правдоподобия со следующего шага. Как и раньше, прогнозируемая переменная r_i принимает значение 1 или 0.

Обозначим как $p(r_i = 1|X = x)$ (условную) вероятность того, что $r_i = 1$ при заданном значении фактора/предиктора.

Возьмем определенную реализацию r_i. Вероятность ее наблюдения можно описать следующей формулой:

$$p(r_i = 1|X = x)^{r_i}(1 - p(r_i = 1|X = x))^{1-r_i}$$

Степени в этом уравнении работают как 'переключатели'. В зависимости от наблюдаемого значения r_i - 0 или 1, 'загорается/включается' та или иная вероятность.

Метод максимального правдоподобия

Допустим, что наблюдаемые значения r_i независимы. Чему равна вероятность наблюдения комбинации из двух заданных наблюдений?

Вероятность наблюдения первого умноженная на вероятность наблюдения второго. Теперь можно обобщить формулу с 'переключателями' для n наблюдений (всей выборки).

$$l(\beta) = \prod_{i=1}^{n} p(r_i = 1|X = x)^{r_i}(1 - p(r_i = 1|X = x))^{1-r_i}$$

Мы помним, что вероятность наблюдения $r_i = 1$ следующим образом зависит от коэффициентов β:

$$p(r_i = 1|X = x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} = \frac{e^{\beta_0+\beta_1 Factor_1}}{1 + e^{\beta_0+\beta_1 Factor_1}}$$

Теперь наша задача сводится к поиску коэффициентов β, которые бы максимизировали функцию правдоподобия (формально, ее логарифм).

Содержание

Интерпретация логит-модели

В своем стандартном виде логистическую модель тяжело итерпретировать. Для более наглядной интерпретации можно преобразовать уравнение:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{e^{\beta_0+\beta_1 Factor_1}}{1+e^{\beta_0+\beta_1 Factor_1}}$$

в уравнение вида:

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$$

Левая часть уравнение называется шансами наступления определенного события. В нашем случае, шансами того, что наблюдение относится к классу 1. Чем выше такие шансы, тем больше вероятность p(x) превышает вероятность 1 - p(x).

Интерпретация логит-модели

Далее можно осуществить следующее преобразование:

$$log(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 x$$

Левая часть называется лог-шансами или логит. И теперь с точки зрения интерпретации очевидно, что лог-шансы линейно зависят от нашего предиктора/фактора.

Таким образом изменение значения x на одну единицу приведет к изменению лог-шансов на β_1.

В целом, также понятно, какие факторы будут приводить к росту вероятности, а какие ее снижают. Например, если β_1 имеет положительное значение, то увеличение x, будет приводить к росту p(x).

Содержание

Множественная логистическая регрессия

Теперь рассмотрим случай бинарной классификации с использованием нескольких факторов.

По аналогии с простой логистической регрессией оценка вероятности принадлежности наблюдения к заданному классу в модели множественной логистической регрессии описывается следующим уравнением:

$$p(x) = \frac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i Factor_i}}{1 + e^{\beta_0 + \sum_{i=1}^{n} \beta_i Factor_i}}$$

Соотвественно, логит будет выглядеть следующим образом:

$$log(\frac{p(x)}{1 - p(x)}) = \beta_0 + \sum_{i=1}^{n} \beta_i Factor_i$$

И здесь, как и раньше, наша задача сводится к оценке оптимальных коэффициентов бета с помощью метода максимального правдоподобия.

Содержание

Логистическая регрессия для m-классов

Перейдем к случаю, когда у нас имеется не два класса (бинарная классификация), а неколько (m) классов. Пусть, как и прежде, у нас есть множество факторов/предикторов. Тогда задача классификации сводится к поиску функции $f : \mathbb{R}^n \to [1..., m]$.

Пусть x вектор содержащий значения факторов/предикторов. Для случая n предикторов он будет иметь размерность n+1: первый элемент этого вектора равен единице, остальные - содержат значения соотвествующего фактора.

Обозначим за $p(Y = k|X = x)$ вероятность (условную) того, что наблюдение относится к классу k при условии, что вектор предикторов равен x.

Как и раньше мы будем оперировать понятиями логит и вероятности принадлежности наблюдения к заданному классу. Единственно, поскольку классов у нас теперь несколько (m), то логит и оценки вероятностей будут базироваться на системе из m-1 уравнений.

Логистическая регрессия для m-классов

Соотвественно, логит (по аналогии с тем, что мы видели раньше) будет описываться системой следующих уравнений:

$$log(\frac{p(Y=1|X=x)}{P(Y=m|X=x)}) = x^T\beta_1 = \beta_{01} + \sum_{i=1}^{n}\beta_{i1}Factor_{i1}$$

$$log(\frac{p(Y=2|X=x)}{P(Y=m|X=x)}) = x^T\beta_2 = \beta_{02} + \sum_{i=1}^{n}\beta_{i2}Factor_{i2}$$

...

$$log(\frac{p(Y=m-1|X=x)}{P(Y=m|X=x)}) = x^T\beta_{m-1} = \beta_{0m-1} + \sum_{i=1}^{n}\beta_{im-1}Factor_{im-1}$$

Логистическая регрессия для m-классов

Оценка вероятности принадлежности наблюдения к заданному классу будет описываться следующей системой из m уравнений:

$$p(Y = k|X = x) = \frac{e^{\beta_{0k} + \sum_{i=1}^{n} \beta_{ik} Factor_i}}{1 + \sum_{k=1}^{m-1} e^{(\beta_{0k} + \sum_{i=1}^{n} \beta_{ik} Factor_i)}}, k = 1, 2...m-1$$

$$p(Y = m|X = x) = \frac{1}{1 + \sum_{k=1}^{m-1} e^{(\beta_{0k} + \sum_{i=1}^{n} \beta_{ik} Factor_i)}}$$

При внимательном рассмотрении можно заметить, что в сумме такая система дает 1.

Следующий этап - это конструирование функции правдоподобия для такого случая.

Логистическая регрессия для m-классов

Для этого введем множество бинарных переменных следующего вида: $b_1, b_2, ..., b_k$. Если наблюдаемая величина Y принадлежит классу k, то соотвествующая переменная b_k равна единице, а остальные - нулю.

Зафиксируем определенную реализацию r_i. Вероятность ее наблюдения тогда описывается формулой:

$$p(r_i = k|X = x) = p(r_i = 1|X = x)^{b_1} p(r_i = 2|X = x)^{b_2} ... p(r_i = k|X = x)^{b_k}$$

А функция правдоподобия (при допущении о независимости) равна:

$$l(\beta) = \prod_{i=1}^{n} p(r_i = 1|X = x)^{b_1} p(r_i = 2|X = x)^{b_2} ... p(r_i = k|X = x)^{b_k}$$

Как и ранее для максимизации функции и поиска оптимальных значений коэффициентов β (решение системы уравний методом Ньютона-Рафсона) нам удобнее работать с логарифмом этой функции (функцией лог-правдоподобия).

Содержание

Коэффициент точности и матрица ошибок

Самый простой критерий оценки качества классификации с ипользованием модели - это коэффициент точности (accuracy ratio). Он представляет собой процент правильно классифицированных моделью наблюдений.

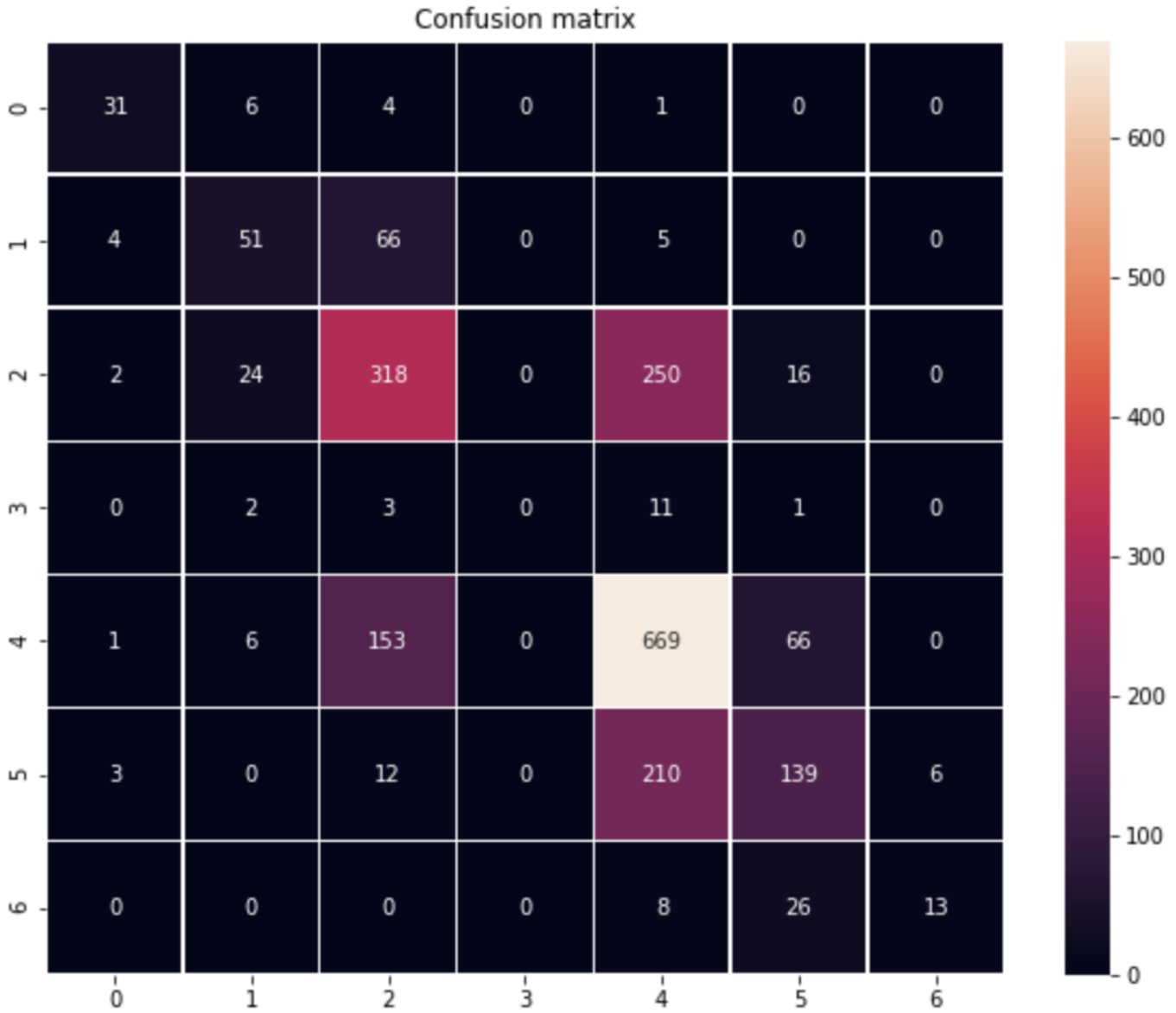
Т.е если accuracy ratio равен 0.8, то модель правильно классифицирует наблюдения в 80 процентах случаев.

Матрица ошибок (confusion matrix) представляет собой распространенный и довольно эффективный способ визуализации качества работы классификационной модели.

Столбцы этой матрицы представляют собой предсказанные классы, а строки - классы, к которым переменная относилась на самом деле. Идеальная модель будет иметь ненулевые значения по главной диагонали и нули - в остальных ячейках.

Давайте проанализируем пример на следующем слайде.

Коэффициент точности и матрица ошибок



Содержание

ROC-кривая и AUCROC

Теперь мы рассмотрим распространенные метрики анализа качества модели для бинарной(!) классификации.

Для начала обозначим наши классы как положительный (1) и отрицательный (0). Теперь рассмотрим систему следующих метрик:

- истинно-положительный результат (true-positive, TP)
- ложно-положительный результат (false-positive, FP)
- истинно-отрицательный результат (true-negative, TN)
- ложно-отрицательный результат (false-negative, FN)

Способность алгоритма предсказывать положительные классы называется чувствительностью модели (или true positive rate, TPR) и вычисляется по формуле:

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

ROC-кривая и AUCROC

Доля ложно-положительных прогнозов (false positive rate) представляет собой количество неверно предсказанных положительных наблюдений к общему количеству отрицательных наблюдений:

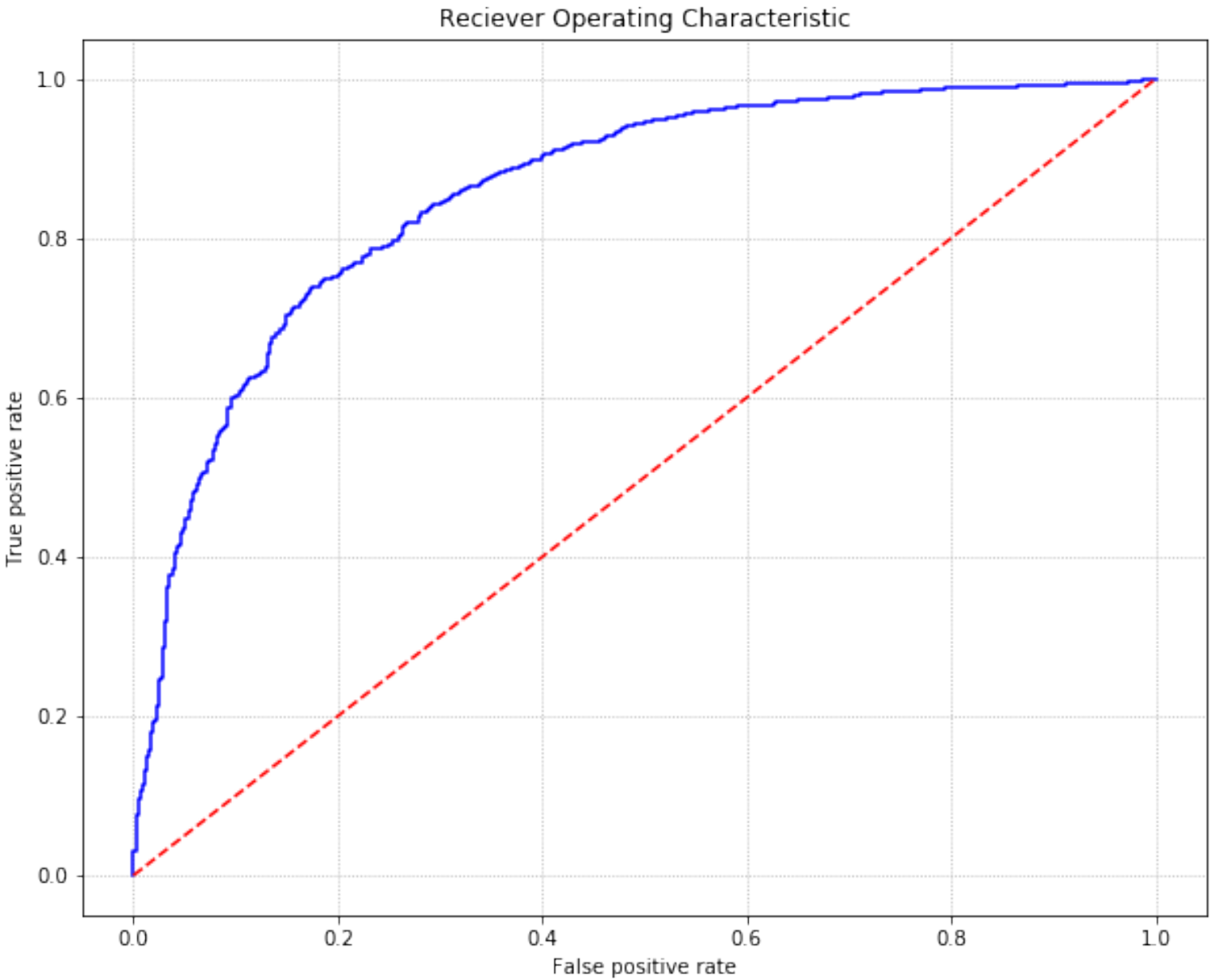
$$FPR = \frac{FP}{FP + TN}$$

Как мы помним, при классификации модель оценивает вероятность того, что наблюдением принадлежит тому (p) или иному (1-p) классу и, с учетом порогового правила (treshold), классифицирует его в соотвествующую категорию.

ROC-кривая представляет собой график отражающий значения TPR и FPR для каждого варианта порогового правила (от 0 до 1).

Показатель AUCROC (area under ROC-curve) дает количественую интерпретацию графика и равен площади под ROC-кривой. Чем выше значение AUCROC, тем лучше (в идеале 1). Значение до 0.5 говорит о том, что модель плохо справляется с классификацией.

ROC-кривая и AUCROC



Литература

Дополнительная литература к сегодняшней лекции:

- Albon, C. (2018) Machine learning with Python cookbook. O'Reilly.
- Hastie, T., Tibshirani, R., Friedman, J. (2017). The elements of statistical learning. Springer. (есть на русском)
- Hosmer, D. (2013) Applied logistic regression. Wiley
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to statistical learning. Springer. (есть на русском)
- Kuhn, M., Johnson, K., (2016). Applied predictive modeling. Springer. (есть на русском)