Лекция 7

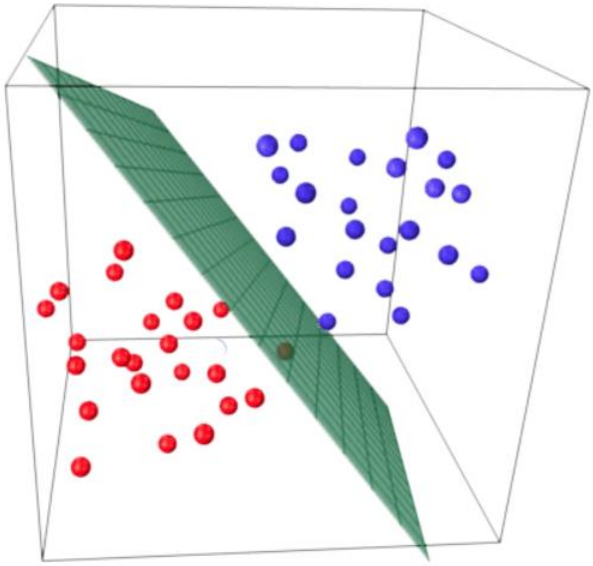
Логистическая регрессия. ROC-анализ

- Линейный классификатор
- Логистическая регрессия
- Ошибки I и II рода
- ROC-кривая, AUC
- Регуляризация

Линейный классификатор

Основная идея линейного классификатора заключается в том, что признаковое пространство может быть разделено гиперплоскостью на два полупространства, в каждом из которых прогнозируется одно из двух значений целевого класса.

Если это можно сделать без ошибок, то обучающая выборка называется линейно разделимой.



Рассмотрим задачу бинарной классификации, причем метки целевого класса обозначим "+1" (положительные примеры) и "-1" (отрицательные примеры).

Один из самых простых линейных классификаторов получается на основе регрессии вот таким образом:

$$a(\vec{x}) = sign(\vec{w}^T x),$$

Где

- $\vec{x}$ – вектор признаков примера (вместе с единицей);
- $\vec{w}$ – веса в линейной модели (вместе со смещением w_0);
- $sign(\bullet)$ – функция "сигнум", возвращающая знак своего аргумента;
- $a(\vec{x})$ – ответ классификатора на примере $\vec{x}$.

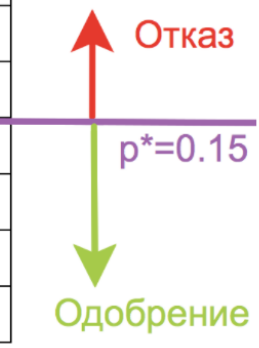
Логистическая регрессия

Логистическая регрессия является частным случаем линейного классификатора, но она обладает хорошим "умением" – **прогнозировать вероятность** p+ отнесения примера $\vec{x}_i$ к классу "+":

$$p_+ = P\left(y_i = 1 \mid \vec{x}_i, \vec{w}\right)$$

Прогнозирование не просто ответа ("+1" или "-1"), а именно вероятности отнесения к классу "+1" во многих задачах является очень важным бизнес-требованием. Например, в задаче *кредитного скоринга*, где традиционно применяется логистическая регрессия, часто прогнозируют вероятность невозврата кредита (p+). Клиентов, обратившихся за кредитом, сортируют по этой предсказанной вероятности (по убыванию), и получается *скоркарта* — по сути, рейтинг клиентов от плохих к хорошим.

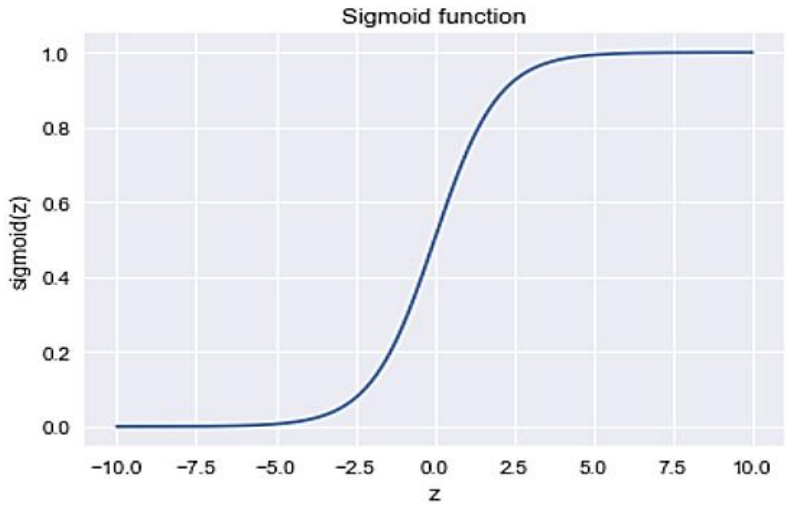
Клиент	Вероятность невозврата
Mike	0.78
Jack	0.45
Larry	0.13
Kate	0.06
William	0.03
Jessica	0.02



Пример скоркарты: банк выбирает для себя порог p* предсказанной вероятности невозврата кредита (в примере – 0.15) и начиная с этого значения кредит не выдает.

Итак, есть задача - прогнозировать вероятность $p+ \in [0,1]$. Мы уже умеем строить линейный прогноз (линейная регрессия) для $Y \in R$. Каким образом преобразовать полученное значение в вероятность, пределы которой – [0, 1]? Очевидно, для этого нужна некоторая функция $f : R \to [0,1]$. В модели логистической регрессии для этого берется конкретная функция (сигмоида):

$$\sigma(z) = \frac{1}{1 + \exp^{-z}}$$



Обозначим P(X) вероятностью происходящего события X. Тогда отношение вероятностей OR(X) определяется из $\frac{P(X)}{1-P(X)}$, а это — отношение вероятностей того, произойдет ли событие или не произойдет. Очевидно, что вероятность и отношение шансов содержат одинаковую информацию. Но в то время как P(X) находится в пределах от 0 до 1, OR(X) находится в пределах от 0 до ∞.

Если вычислить логарифм OR(X) (то есть называется логарифм шансов, или логарифм отношения вероятностей), то легко заметить, что log OR(X) находится в пределах от - ∞ до ∞. Его-то мы и будем прогнозировать.

Рассмотрим, как логистическая регрессия будет делать прогноз $p_+ = P\left(y_i = 1 \mid \vec{x_i}, \vec{w}\right)$, предположим, что веса w_i известны:

Шаг 1. Вычислить значение $w_0 + w_1 x_1 + w_2 x_2 + \ldots = \vec{w}^T \vec{x}$. (уравнение $\vec{w}^T \vec{x} = 0$ задает гиперплоскость, разделяющую примеры на 2 класса);

Шаг 2. Вычислить логарифм отношения шансов: $\log(OR_+) = \vec{w}^T \vec{x}$.

Шаг 3. Имея прогноз шансов на отнесение к классу "+" – *OR+*, вычислить *p+* с помощью простой зависимости:

$$p_+ = \frac{OR_+}{1 + OR_+} = \frac{\exp^{\vec{w}^T \vec{x}}}{1 + \exp^{\vec{w}^T \vec{x}}} = \frac{1}{1 + \exp^{-\vec{w}^T \vec{x}}} = \sigma(\vec{w}^T \vec{x})$$

Итак, логистическая регрессия прогнозирует вероятность отнесения примера к классу "+" (при условии, что мы знаем его признаки и веса модели) как сигмоид-преобразование линейной комбинации вектора весов модели и вектора признаков примера:

$$p_+(x_i) = P\left(y_i = 1 \mid \vec{x_i}, \vec{w}\right) = \sigma(\vec{w}^T \vec{x_i})$$

Обучить модель, это подобрать такие коэффициенты, при которых логистическая функция потерь будет минимальной.

$$\mathcal{L}_{\updownarrow l\}}(X, \vec{y}, \vec{w}) = \sum_{i=1}^{\ell} \log(1 + \exp^{-y_i \vec{w}^T \vec{x_i}})$$

$\mathcal{L}$ - Это логистическая функция потерь, просуммированная по всем объектам обучающей выборки.

Решение: МНК или метод градиентного спуска

Метрики качества модели

Ошибки I и II рода

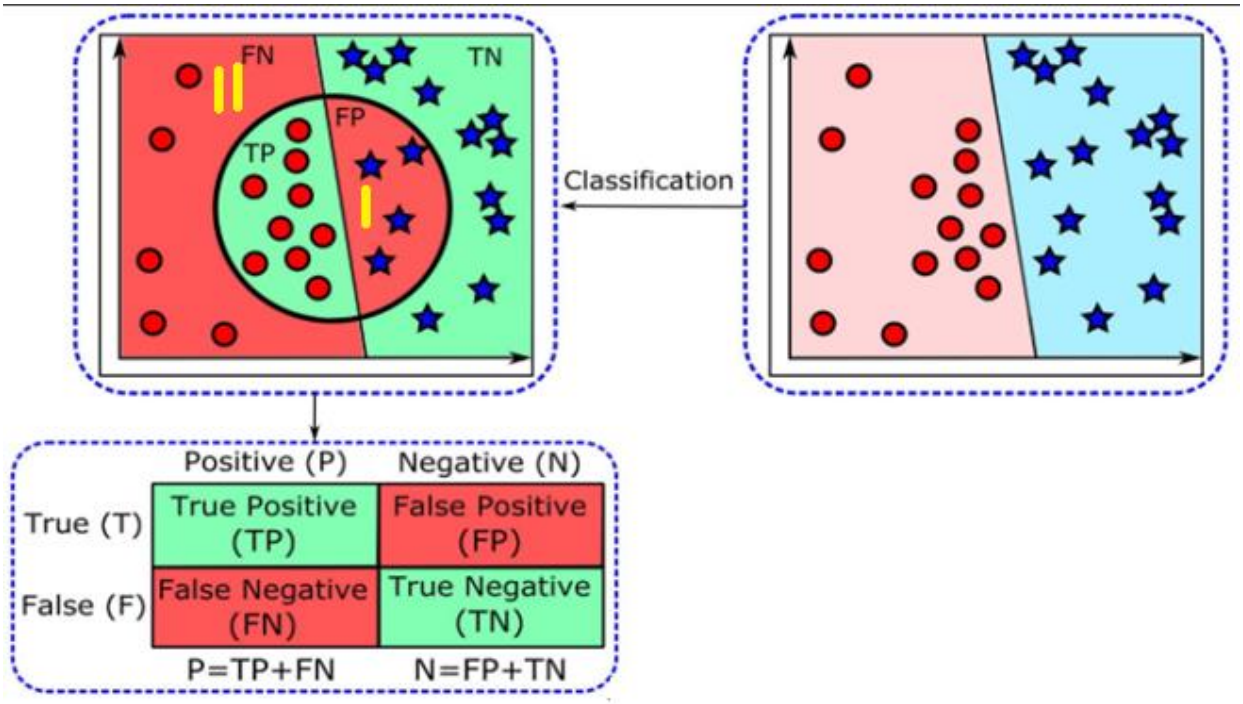
Для понимания сути ошибок I и II рода рассмотрим таблицу сопряженности (confusion matrix), которая строится на основе результатов классификации моделью и фактической (объективной) принадлежностью примеров к классам.

Модель	Фактически положительно	Фактически отрицательно
Положительно	TP	FP
Отрицательно	FN	TN

- TP (**True Positives**) — верно классифицированные положительные примеры (так называемые истинно положительные случаи).
- TN (**True Negatives**) — верно классифицированные отрицательные примеры (истинно отрицательные случаи).
- FN (**False Negatives**) — положительные примеры, классифицированные как отрицательные (ошибка I рода). Это так называемый «ложный пропуск» — когда интересующее нас событие ошибочно не обнаруживается (ложно отрицательные примеры).
- FP (**False Positives**) — отрицательные примеры, классифицированные как положительные (ошибка II рода). Это ложное обнаружение, т.к. при отсутствии события ошибочно выносится решение о его присутствии (ложно положительные случаи).

Ошибка I-го рода (FP): чужого приняли за своего

Ошибка II-го рода (FN): своего приняли за чужого



Что является положительным событием, а что — отрицательным, зависит от конкретной задачи. *Например, если мы прогнозируем вероятность наличия заболевания, то положительным исходом будет класс «Больной пациент», отрицательным — «Здоровый пациент». И наоборот, если мы хотим определить вероятность того, что человек здоров, то положительным исходом будет класс «Здоровый пациент», и так далее.*

При анализе чаще оперируют не абсолютными показателями, а относительными — долями (rates), выраженными в процентах:

- Доля истинно положительных примеров (**True Positives Rate**):

$$TPR = \frac{TP}{TP + FN} \cdot 100\,\%$$

- Доля ложно положительных примеров (**False Positives Rate**):

$$FPR = \frac{FP}{TN + FP} \cdot 100\,\%$$

Введем еще два определения: **чувствительность** и **специфичность модели**. Ими определяется объективная ценность любого бинарного классификатора.

Чувствительность (Sensitivity) —доля истинно положительных случаев:

$$S_e = TPR = \frac{TP}{TP + FN} \cdot 100\,\%$$

Специфичность (Specificity) — доля истинно отрицательных случаев, которые были правильно идентифицированы моделью:

$$S_p = \frac{TN}{TN + FP} \cdot 100\,\%$$

Модель с высокой чувствительностью часто дает истинный результат при наличии положительного исхода (обнаруживает положительные примеры). Наоборот, **модель с высокой специфичностью чаще дает истинный результат при наличии отрицательного исхода** (обнаруживает отрицательные примеры). Если рассуждать в терминах медицины — задачи диагностики заболевания, где модель классификации пациентов на больных и здоровых называется диагностическим тестом, то получится следующее:

- ✓ **Чувствительный диагностический тест проявляется в гипердиагностике** — максимальном предотвращении пропуска больных.
- ✓ **Специфичный диагностический тест диагностирует только доподлинно больных.** Это важно в случае, когда, например, лечение больного связано с серьезными побочными эффектами и гипердиагностика пациентов не желательна.

ROC-анализ

ROC-кривая (Receiver Operator Characteristic) часто используется для представления результатов бинарной классификации в машинном обучении. Название пришло из систем обработки сигналов. Поскольку классов два, один из них называется классом с положительными исходами, второй — с отрицательными исходами. **ROC-кривая показывает зависимость количества верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров**.

В терминологии ROC-анализа первые называются истинно положительным, вторые — ложно отрицательным множеством. При этом предполагается, что **у классификатора имеется некоторый параметр, варьируя который, мы будем получать то или иное разбиение на два класса. Этот параметр часто называют порогом,** или точкой отсечения (cut-off value). В зависимости от него будут получаться различные величины ошибок I и II рода.

В логистической регрессии порог отсечения изменяется от 0 до 1 — это и есть расчетное значение уравнения регрессии. Будем называть его рейтингом.

ROC-кривая получается следующим образом:

- ✓ Для каждого значения порога отсечения, которое меняется от 0 до 1 с шагом d_x (например, 0.01) рассчитываются значения чувствительности Se и специфичности Sp. В качестве альтернативы порогом может являться каждое последующее значение примера в выборке.
- ✓ Строится график зависимости: по оси Y откладывается чувствительность Se, по оси X — $FPR = 100 - Sp$ — доля ложно положительных случаев.

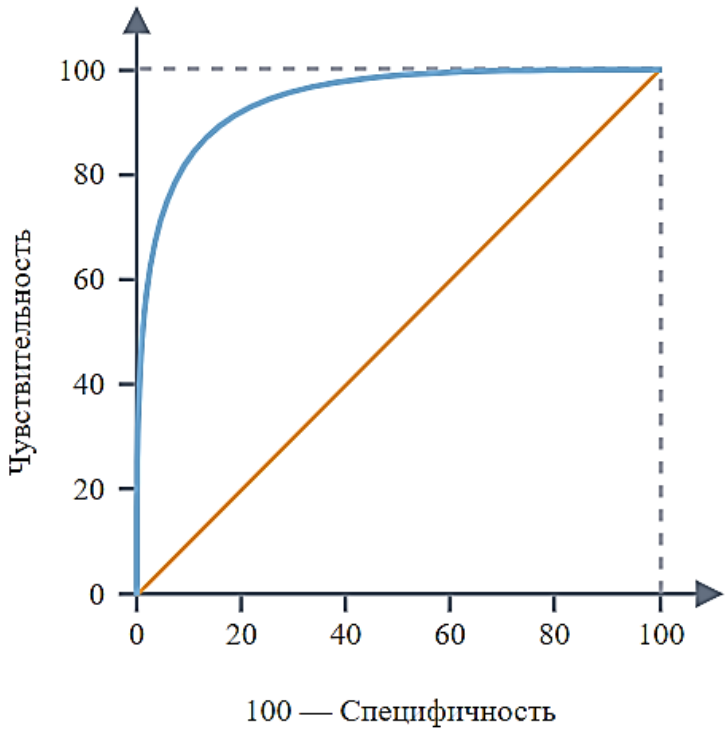


Рис. 2 — ROC-кривая

Для идеального классификатора график ROC-кривой проходит через верхний левый угол, где доля истинно положительных случаев составляет 100% или 1,0 (идеальная чувствительность), а доля ложно положительных примеров равна нулю. Поэтому чем ближе кривая к верхнему левому углу, тем выше предсказательная способность модели. Наоборот, чем меньше изгиб кривой и чем ближе она расположена к диагональной прямой, тем менее эффективна модель. Диагональная линия (y=x) соответствует «бесполезному» классификатору, т.е. полной неразличимости двух классов.

При визуальной оценке ROC-кривых расположение их относительно друг друга указывает на их сравнительную эффективность. Кривая, расположенная выше и левее, свидетельствует о большей предсказательной способности модели. Так, на рис. 3 видно, что модель «A» лучше.

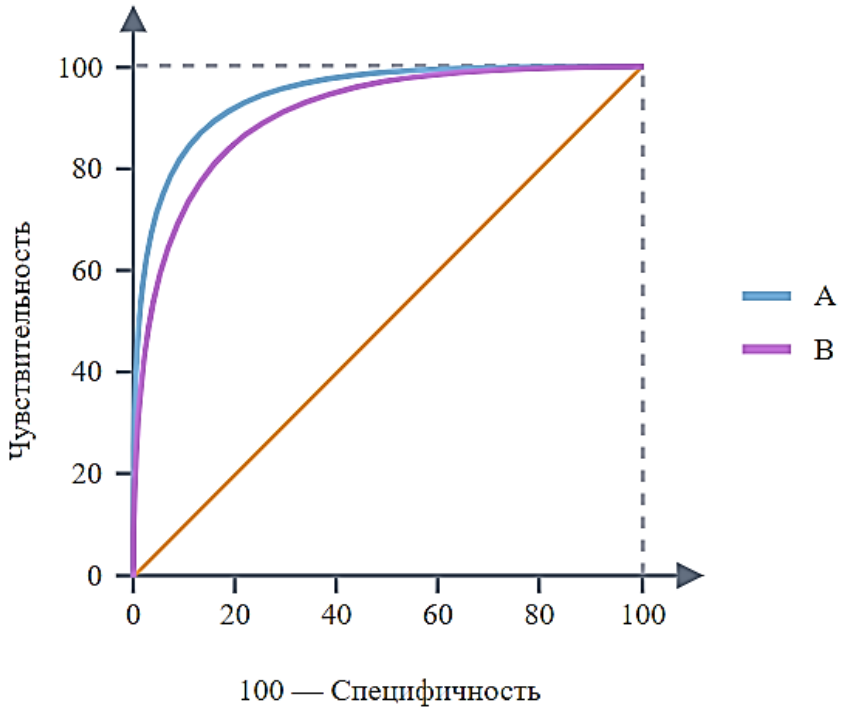


Рис. 3 — Сравнение ROC-кривых

Визуальное сравнение кривых ROC не всегда позволяет выявить наиболее эффективную модель. Своеобразным методом сравнения ROC-кривых является **оценка площади под кривыми**. Теоретически она изменяется от 0 до 1,0, но, поскольку модель всегда характеризуются кривой, расположенной выше положительной диагонали, то обычно говорят об изменениях от 0,5 («бесполезный» классификатор) до 1,0 («идеальная» модель).

Эта оценка может быть получена непосредственно вычислением площади под многогранником, ограниченным справа и снизу осями координат и слева вверху — экспериментально полученными точками (рис. 4). Численный показатель площади под кривой называется **AUC** (Area Under Curve).

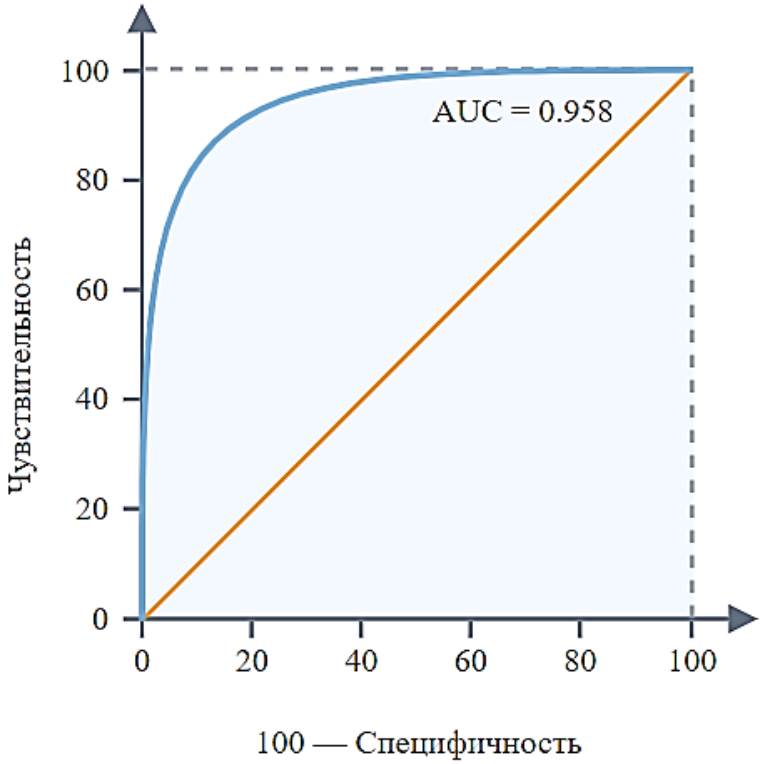


Рис. 4 — Площадь под ROC-кривой

С большими допущениями можно считать, что чем больше показатель AUC, тем лучшей прогностической силой обладает модель. Однако следует знать, что:

- показатель AUC предназначен скорее для сравнительного анализа нескольких моделей;
- AUC не содержит никакой информации о чувствительности и специфичности модели.

В литературе иногда приводится следующая экспертная шкала для значений AUC, по которой можно судить о качестве модели:

Интервал AUC	Качество модели
0,9-1,0	Отличное
0,8-0,9	Очень хорошее
0,7-0,8	Хорошее
0,6-0,7	Среднее
0,5-0,6	Неудовлетворительное

Идеальная модель обладает 100% чувствительностью и специфичностью. Однако на практике добиться этого невозможно, более того, невозможно одновременно повысить и чувствительность, и специфичность модели. Компромисс находится с помощью порога отсечения, т.к. пороговое значение влияет на соотношение Se и Sp. Можно говорить о **задаче нахождения оптимального порога отсечения** (optimal cut-off value).

Порог отсечения нужен для того, чтобы применять модель на практике: относить новые примеры к одному из двух классов. **Для определения оптимального порога нужно задать критерий его определения**, т.к. в разных задачах присутствует своя оптимальная стратегия. Критериями выбора порога отсечения могут выступать:

- ✓ Требование минимальной величины чувствительности (специфичности) модели. Например, нужно обеспечить чувствительность теста не менее 80%. В этом случае оптимальным порогом будет максимальная специфичность (чувствительность), которая достигается при 80%.
- ✓ Требование максимальной суммарной чувствительности и специфичности модели, т.е.

$$Cutt\_off_o = \max_k(Se_k + Sp_k)$$

В этом случае значение порога обычно предлагается пользователю по умолчанию.

- ✓ Требование баланса между чувствительностью и специфичностью, т.е. когда Se ≈ Sp

$$Cutt\_off_o = \min_k |Se_k - Sp_k|$$

В этом случае порог есть точка пересечения двух кривых, когда по оси X откладывается порог отсечения, а по оси Y — чувствительность или специфичность модели (рис. 5).

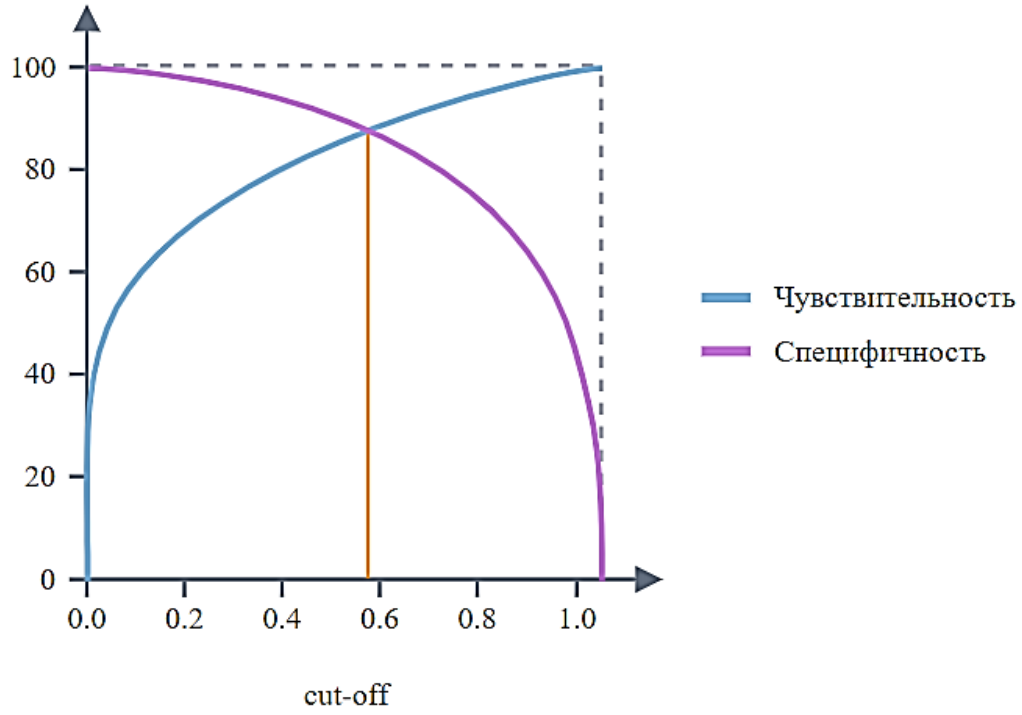
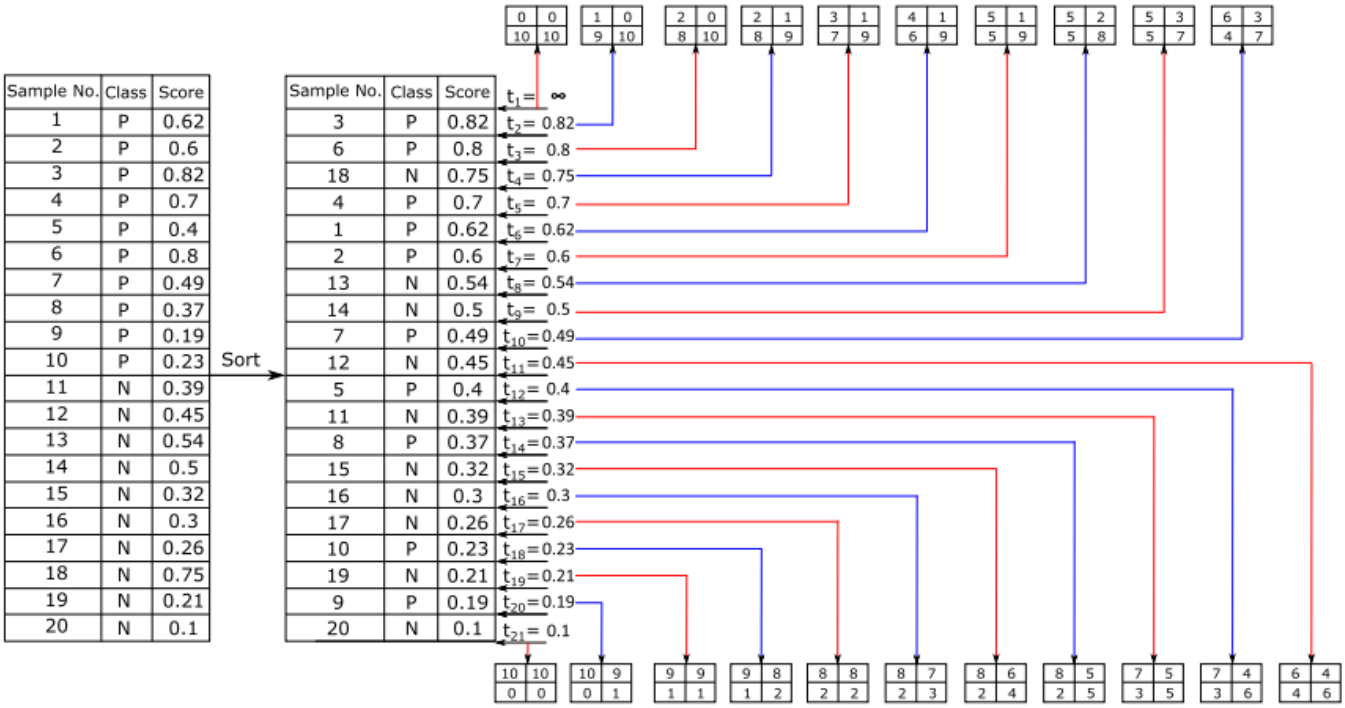
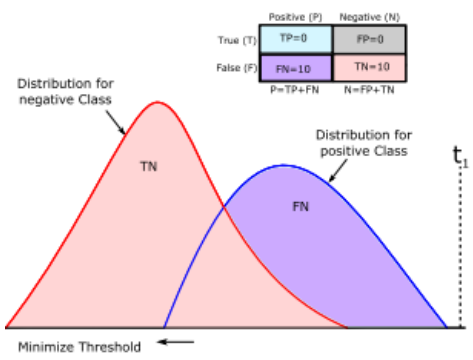


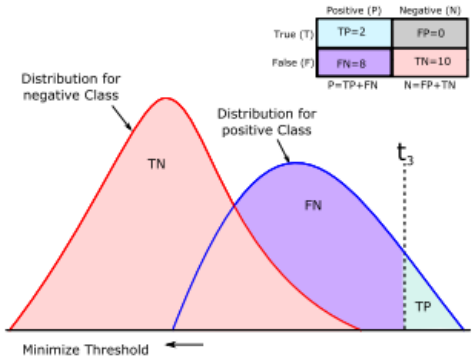
Рис. 5 — «Точка баланса» между чувствительностью и специфичностью

Наглядный пример расчета TPR и FPR при изменении порогового значения.

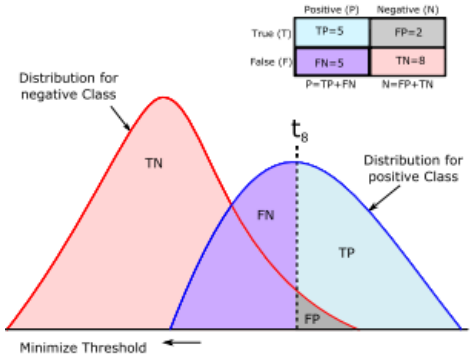




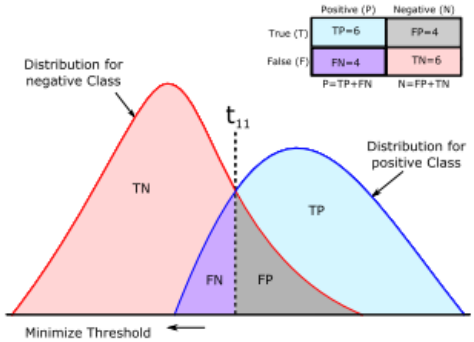
(a)



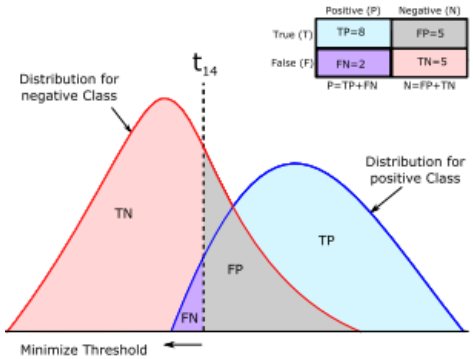
(b)

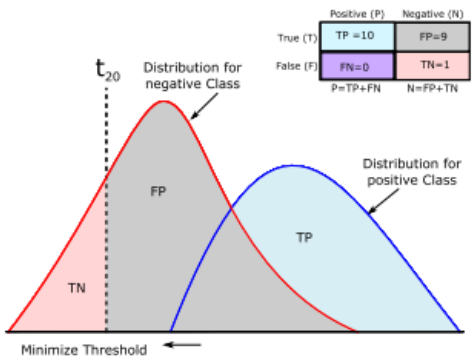


(c)



(d)





https://www.researchgate.net/publication/327148996_Classification_Assessment_Methods_a_detailed_tutorial

Регуляризация

Регуляризация в статистике, машинном обучении— метод добавления некоторых дополнительных ограничений к условию с целью решить некорректно поставленную задачу или предотвратить переобучение. Эта информация часто имеет вид штрафа за сложность модели.

Переобучение в большинстве случаев проявляется в том, что в получающихся многочленах слишком большие коэффициенты. Соответственно, необходимо добавить в целевую функцию штраф за слишком большие коэффициенты.

Некоторые виды регуляризации:

- L_1-регуляризация (англ. *lasso regression*), или **регуляризация**

$$L_1 = \sum_i (y_i - y(t_i))^2 + \lambda \sum_i |a_i|.$$

- L_2- **регуляризация**, или **регуляризация Тихонова** (в англоязы

уравнений позволяет балансировать между соответствием дан

$$L_2 = \sum_i (y_i - y(t_i))^2 + \lambda \sum_i {a_i}^2.$$

sklearn.linear_model.LogisticRegression

class sklearn.linear_model.LogisticRegression(*penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None*)

Этот класс реализует регуляризованную логистическую регрессию с использованием решателей newton-cg, sag, saga и lbfgs.

По умолчанию применяется регуляризация.

Мультиклассовая логистическая регрессия

по умолчанию для solver установлено значение " liblinear ", которое не поддерживает мультикласс. Мультикласс поддерживается в 'newton-cg,' lbfgs, 'sag,' saga