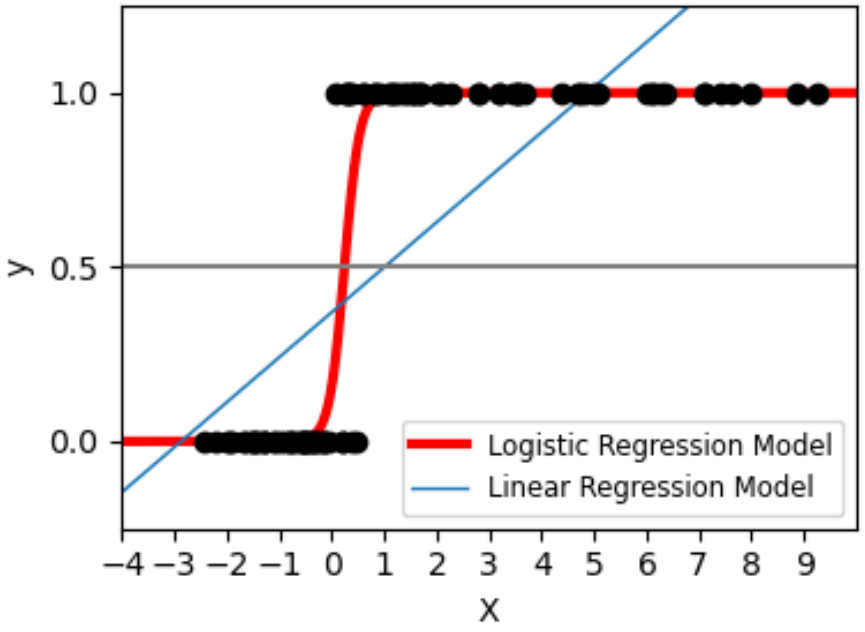
Классификационные модели и методы оценки их дискриминационной способности. Логистической Регрессии.

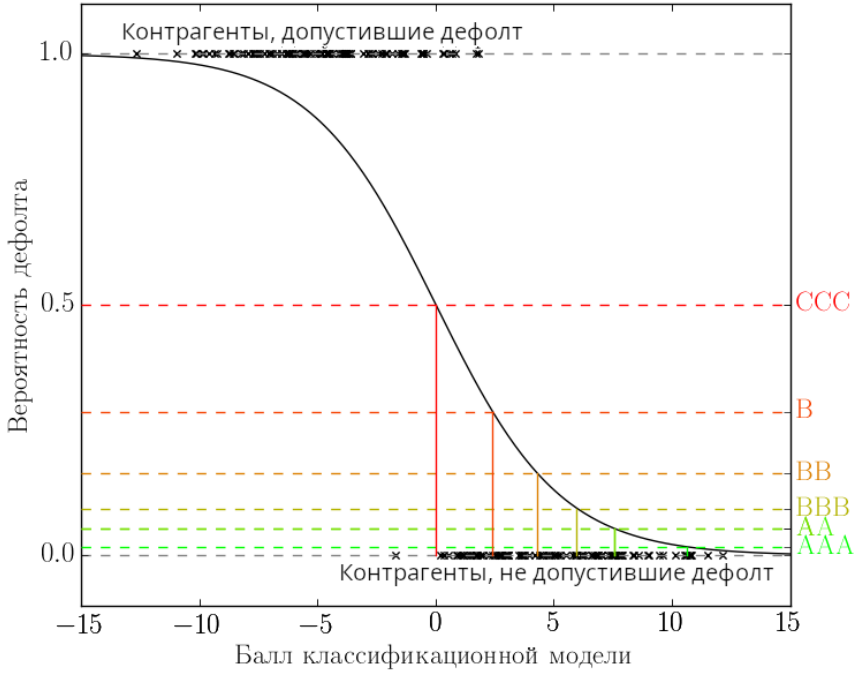
План вебинара

- Класификационные модели бинарной классификации. Классификация при помощи Линейной регрессии.
- LogLoss. Логистическая регрессия и её уравнения.
- Матрица ошибок. Точность классификационных моделей.
- Методы оценки дискриминационной способности класификационной модели.

Почему логистическая регрессия называется «регрессией»?

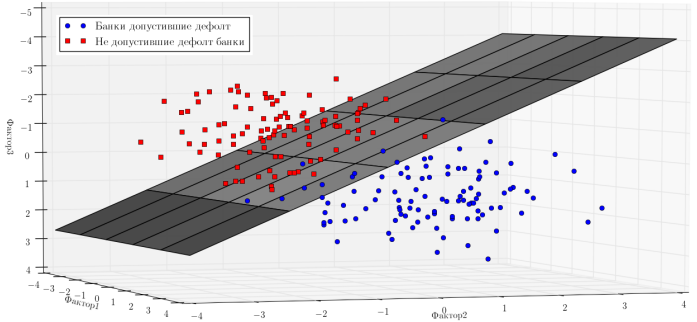


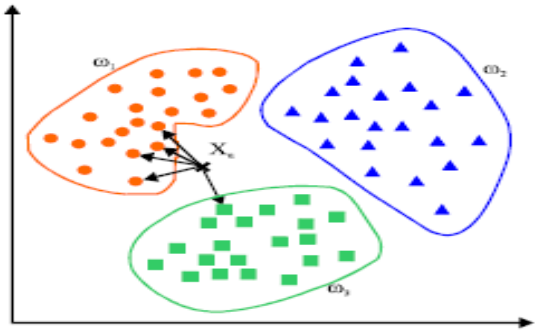
Взято с
https://scikit-learn.org/stable/auto_examples/linear_model/plot_logistic.html#sphx-glr-auto-examples-linear-model-plot-logistic-py



Место логистической регрессии среди классификаторов

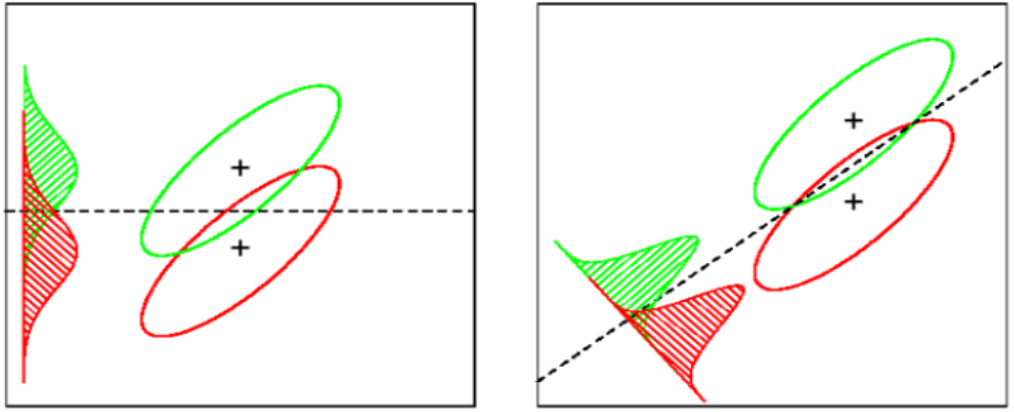
- Существуют методы, ориентирующиеся на глобальные особенности распределения данных(например SVM) и на локальные(например KNN)
- Логистическая регрессия ориентируется на глобальные а не локальные особенности распределения данных





Взято с https://www.mathworks.com/matlabcentral/fileexchange/63621-knn-classifier

Место логистической регрессии среди классификаторов



Взято с https://www.dbs.ifi.lmu.de/Lehre/MaschLernen/SS2016/Skript/LinearClassifiers2016.pdf

- Будучи линейной моделью, логистическая регрессия проявляет значимую дискриминационную способность при наличии дискриминационной способности у факторов модели.
- Не работает на факторах, показательных только в соотношениях друг с другом. Например: состояние отрасли и позиция в отрасли контрагента

Подготовка данных для логистической регрессии

- Убеждаемся что факторы по отдельности оказывают ожидаемое влияние, т. е. для больших/меньших значений фактора большая/меньшая принадлежность к определенному классу.
- Если есть дискретный фактор, значения которого равнозначны, разбиваем его на несколько бинарных факторов, соответствующих каждому значению.
- Если имеется аналогичная ситуация с непрерывным фактором, используем набор бинарных факторов показывающих принадлежность значения исходного фактора к диапазону.
- Если есть два фактора влияющих вместе, а не по отдельности, комбинируем их в один фактор, используя их соотношение или иным образом. Например отношение размера займа к доходу.
- Следим чтобы у всех факторов были одинаковые диапазоны. В случае непрерывных факторов не лишним будет привести к одному распределению.
- Поскольку Логистическая Регрессия- линейная модель, необходимо следить за отсутствием значимых корреляций и мультиколлинеарности среди факторов.

Log-Loss и максимизация ожидания

- Вероятность наблюдать исходные данные: $p(d|x)=\prod_i (p_i*d_i+(1-p_i)*(1-d_i))$
  - p_i вероятность принадлежность к целевому классу
  - d_i флаг принадлежности(1 в случае принадлежности к целевому классу, 0 в противном случае)
- Если вы полнить логарифмирование, то произведение заменится на сумму:
  $\log(p(d|x))=\sum_i \log(p_i*d_i+(1-p_i)*(1-d_i))=\sum_i (d_i*\log(p_i)+(1-d_i)*\log(1-p_i))$
- $LogLoss=-\frac{1}{N}\sum_i (d_i*\log(p_i)+(1-d_i)*\log(1-p_i))$ Принимает значения от нуля(вероятности равны нулям и единицам и расставлены правильно) до бесконечности

Основное уравнение логистической регрессии

- Для Логистической Регрессии используется $p(score)=\frac{1}{1+\exp(-score)}$
- У этой формулы есть свойство $1-p(score)=p(-score)$
- Воспользовавшись предыдущей формулой и тем что члены с d_i и $1-d_i$ взаимоисключают друг друга получим для Логистической Регрессии:

  $LogLoss=\frac{1}{N}\sum \log(1+\exp(-y_i*score_i))$ где y_i равен 1 или -1 в зависимости от класса
- Собственно $score(x)=a_0+\sum_i a_i x_i$, где коэффициенты a подлежат определению в ходе максимизации, x_i значения факторов

Нахождение производной для LogLoss

- Для производной LogLoss
$$\frac{\partial LogLoss}{\partial w_j} = \sum_i -y_i \frac{\exp(-y_i score_i)}{1+\exp(-y_i score_i)} \frac{\partial score_i}{\partial w_j} = \sum_i \frac{-y_i}{1+\exp(y_i score_i)} \frac{\partial score_i}{\partial w_j}$$
- Для коэффициента при факторе $\frac{\partial score_i}{\partial w_i} = x_i$, для свободного коэффициента $\frac{\partial score_i}{\partial w_j} = 1$.
- Если записать в виде строки Z, где $Z_i = \frac{-y_i}{1+\exp(y_i score_i)}$, то выражение для коэффициентов при факторах запишется в виде матричного перемножения $Z * X$, а для свободного коэффициента как сумма $\sum_i Z_i$

Монотонная классификация логистической регрессией

- Монотонная классификация- классификация при которой один или несколько коэффициентов должны быть неотрицательны.
- Для коэффициентов которые должны быть неотрицательны возможно определить веса:

$$w_i = \frac{a_i}{\sum_{j \in RequirePositive} a_j}$$

Регулярязационная поправка в логистической регрессии

- Используется для избежания оверфитинга: запоминания обучающего набора данных с потерей способности к обобщению.
- Используется как поправка к штрафной функции. Добавляется с регулярязационным коэффициентом C : $LogLoss + C * Reg$
- Бывает двух видов:
  - $L_1 : \sum_i |a_i|$
  - $L_2 : \sum_i a_i^2$
- В случае полностью монотонной классификации для L_2 имеем $Reg \sim \sum_i w_i^2 = HHI(w)$, где HHI индекс Херфиндаля-Хиршмана, который минимален при равномерном распределении весов.
- Таким образом регуляризация L_1 минимизирует количество/влияние используемых факторов, а регуляризация L_2 влияние используемых факторов усредняет.

Мультиклассовая классификация

- В случае задачи мультиклассовой классификации можно построить несколько логистических регрессий по принципу свой класс против остальных, а можно использовать софтмакс-классификатор.
- В софтмакс классификаторе строится несколько линейных комбинаций для рассчетов баллов, соответствующих определенным классам. Вероятность принадлежности к классу i определяется по формуле $p_i = \frac{\exp(score_i)}{\sum_j \exp(score_j)}$
- Софтмакс-классификатор в случае двух классов превращается в логистическую регрессию.

Оценка дискриминационной способности

	Для положительного класс	Для отрицательного класса
Предсказан пололжительный класс	Правильно(TP)	Ошибка первого рода(FP)
Предсказан отрицательный класс	Ошибка второго рода(FN)	Правильно(TN)

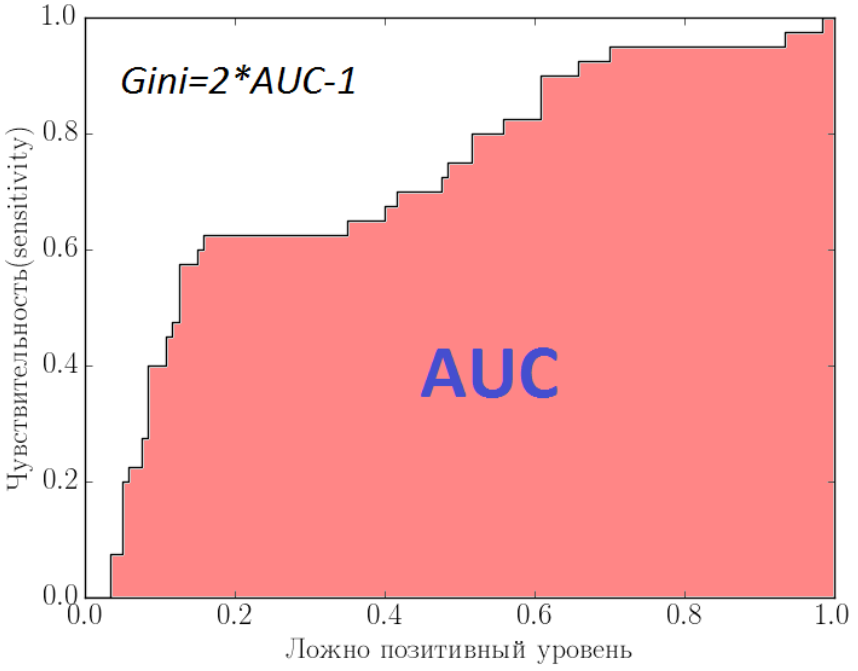
- Ошибки первого и второго рода взаимосвязаны(чем меньше одна, тем больше другая) и зависят от выбранного порогового значения балла.

- Дискриминационная способность- способность к разделению по значению классификационного бала на классы, соответствующие заложенным в модель.
- Оценка дискриминационной способности подразумевает оценку значимости или оценку доли ошибок при классификации.
- Существует два семейства подходов к оценке дискриминационной способности:
  - Учитывающих правильные негативные предсказания(TN): ROC, KS
  - Игнорирующая TN: Precision-Recall

Receive Operating Characteristic

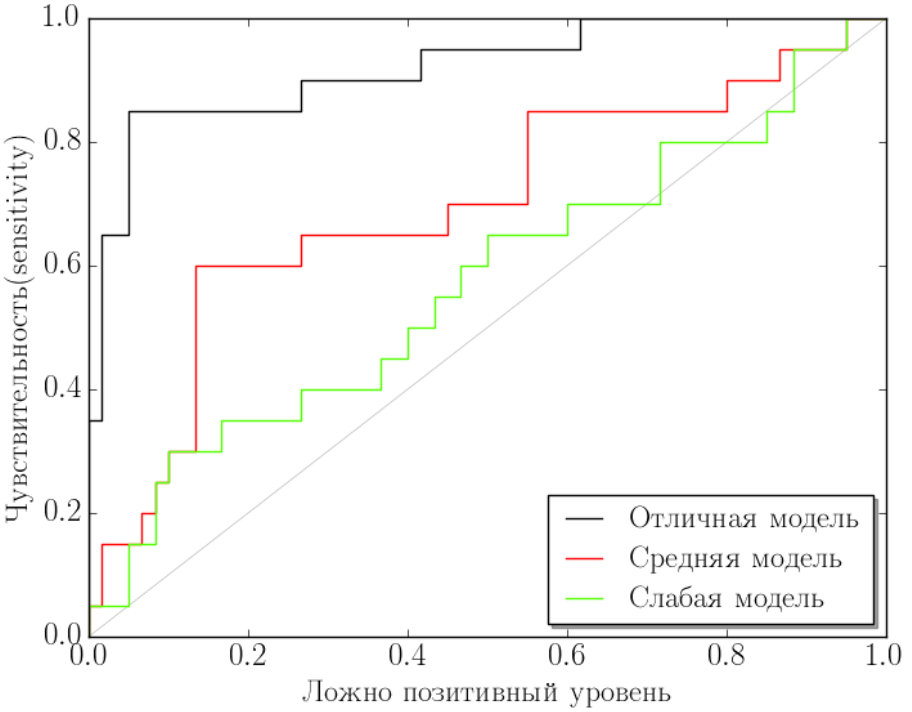
	Для положительного класс	Для отрицательного класса
Предсказан пололжительный класс	Правильно(TP)	Ошибка первого рода(FP)
Предсказан отрицательный класс	Ошибка второго рода(FN)	Правильно(TN)

- Чувствительность (Sensitivity)= TP/(TP+FN)
- Ложно-позитивный уровень (FPR)=FP/(FP+TN)

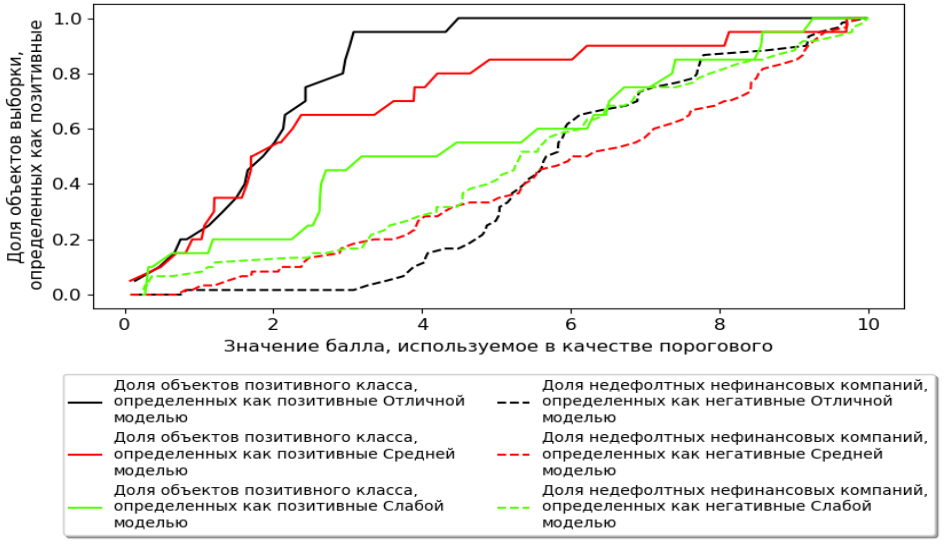


Receive Operating Characteristic

- AUC имеет статистический смысл вероятности того, что у позитивного класса балл больше/меньше, и соответствует U-статистике теста(критерия) Mann-Whitney.
- Также применяется коэффициент Gini принимающий значения от 0 дс 1.



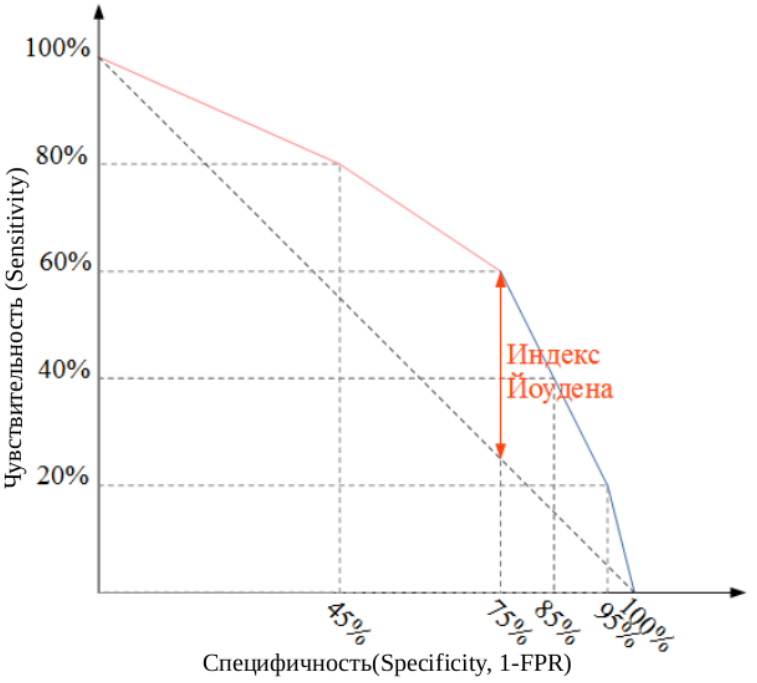
Kolmogorov-Smirnov



- Аналогично ROC, для каждого порогового балла вычисляется доля объектов позитивного и негативного класса, с баллом выше/ниже порогового. Доли равна sensitivity и FPR соответственно.
- Максимальная разность долей равна KS- дистанции Колмогорова-Сммирнова (статистика D из одноименного теста).

ROC и KS сравнение

- Если ROC это интегральная оценка дискриминационной способности, то KS- для значения балла наиболее оптимального для классификационного порога.
- В ROC анализе есть равный KS показатель индекс Йоудена(J)
- KS также выражается через показатель Balanced Accuracy KS=2BA-1,
  - где BA=1/2(sensitivity+specificity),
    - где specificity=TN/(TN+FP)=1-FPR

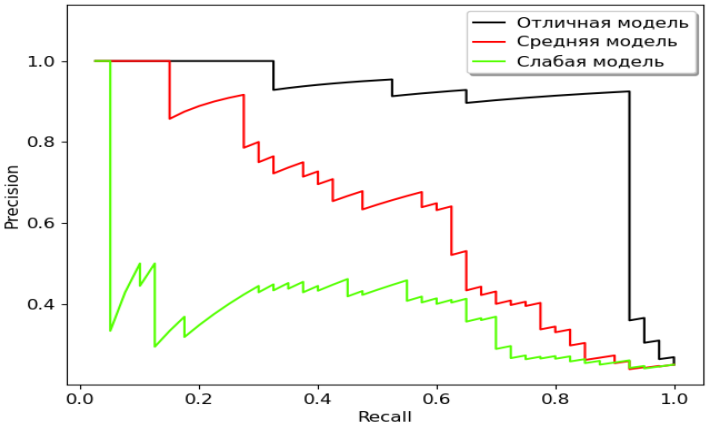


Precision-Recall

	Для положительного класс	Для отрицательного класса
Предсказан пололжительный класс	Правильно(TP)	Ошибка первого рода(FP)
Предсказан отрицательный класс	Ошибка второго рода(FN)	Правильно(TN)

Precision=TP/(TP+FP)

Recall=TP/(TP+FN)



- Если для целей классификации важно правильное определение принадлежности только для одного класса(положительного) используется метрика на основе Precision, Recall.
- Для метрики на основе Precision-Recall можно построить PR-кривые для сравнения классификаторов, а также рассчитать интегральный показатель PR-score.
- В отличие от AUC ROC(AUROC) площадь под кривой PR для расчёта PR-score определяется не при помощи метода трапеций, а более грубо- при помощи метода прямоугольников.
- PR-score может принимать значения от 0 до 1, при этом случайному классификатору соответствует значение равное доле объектов положительного класса в выборке.
- В случае фиксированного порога можно использовать F1-score равный $2\frac{Precision * Recall}{Precision + Recall}$