МЕТРИКИ МАШИННОГО ОБУЧЕНИЯ

Бинарная классификация

Введение

Метрики бинарных классификаторов пришли из теории связи, в задаче бинарного различения или обнаружения импульсов. Так как коррелятор или согласованный фильтр - первое устройство, которое решало эту задачу, такие метрики были сначала использованы именно для этого. Сейчас эти метрики популярны в задачах классификации машинного обучения.

Confusion matrix

Матрица ошибок или confusion matrix - таблица, отображающая качество алгоритма.

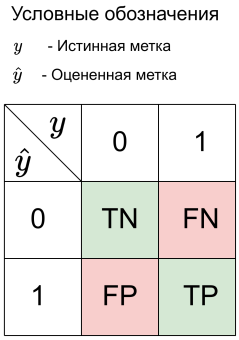

Figure 1: Confusion matrix

По горизонтали отмечены истинные метки класса, по вертикали их оценки. Внутри матрицы указаны подсчитанные значения. Где:

1. TN - True negative - верноотрицательная оценка метки класса. То есть метка равна 0 и оценка равна 0
2. TP - True positive - верноположительная оценка метки класса. То есть метка равна 1 и оценка равна 1
3. FN - False negative - ошибочно отрицательная оценка метки класса. То есть метка равна 0 а принято решение 1
4. FP - False positive - ошибочноотрицательная оценка метки класса. То есть метка равна 1, а принято решение 0

Accuracy(Доля правильных ответов)

Accuracy или точность - метрика которая показывает долю верно определенных классов или оценку вероятности прогноза вероопределенного класса.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Такая метка актуальна для равновероятных классов, но неакутальна для несбалансированных. Пример:

- TP=5, FN=5, TN=90, FP=10. Тогда $Acc = \frac{TP+TN}{TP+TN+FP+FN} = \frac{5+90}{5+90+10+5} = \frac{95}{110}$

Однако если мы все письма будет предсказывать как не спам, мы получим другую метрику:

- TP=0, FN=10, TN=100, FP=0. Тогда $Acc = \frac{TP+TN}{TP+TN+FP+FN} = \frac{0+100}{0+100+0+10} = \frac{100}{110}$

Качество прогноза уменьшилов, однако значение метрики увеличилось. Это показывает то, что accuracy - метрика для классификации равновероятных классов

Precision(точность)

Precision другая метка, которая решает проблему несбалансированных классов. Уравнение этой метрики следующее:

$$P = \frac{TP}{TP + FP} \tag{2}$$

Такая метрика по сути вещей показывает способность отличать данный класс от других.

Recall(полнота)

Recall - другая метрика, которая используется в тех случаях, когда потеря правильного ответа для класса существенна. Например, в сутки делается много тысяч звонков, необходимо определить те их них, которые являются мошенническими. Если мы будем, классифицировать мошеннические действия как немошеннические(FN), это будет дорогого стоить.

$$R = \frac{TP}{TP + FN} \tag{3}$$

Отсюда следует, что метрика Recall показывает возможность определять данный класс.

F-мера

F-мера - метрика, вычисляющая среднее геометрическое между метриками Precision и Recall. Среднее геометрическое это:

$$HM = \frac{n}{\sum_{i=0}^{n} \frac{1}{x_i}}$$

Таким образом, F-Мера имеет значение:

$$F - measure = \frac{2}{\frac{1}{R} + \frac{1}{P}} = \frac{2}{\frac{P}{RP} + \frac{R}{RP}} = \frac{2RP}{P + R} \tag{4}$$

Взвешивая метрики precision и recall коэффициентами $\alpha = \frac{1}{(\beta^2+1)}$ и $1 - \alpha$, так чтобы суммарный вес был равен 1 можно получить взвешенную F-Меру

$$F - measure = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{1}{\frac{\alpha}{P} + \frac{1}{R} - \frac{\alpha}{R}} = \frac{1}{\frac{R\alpha}{RP} + \frac{P}{RP} - \frac{P\alpha}{RP}} = \frac{RP}{R\frac{1}{(\beta^2+1)} + P\frac{(\beta^2+1)-1}{(\beta^2+1)}} = \left(\beta^2+1\right)\frac{RP}{\beta^2 P + R}$$

Если $\left(\beta^2+1\right) = 2$ то метрика превращается в обычную F-Меру. Коэффициент α позволяет определить степень важности Recall и Precision метрики

ROC-AUC

ROC-Reciever operating characteristics - характеристика пришедшая из задач связи, в которых требуется подобрать оптимальный порого. ROC кривая показывает зависимость TPR(FPR) при различных значениях порога. AUC - Area under curve - площадь под ROC кривой - характеристика, показывающая качество классификатора.

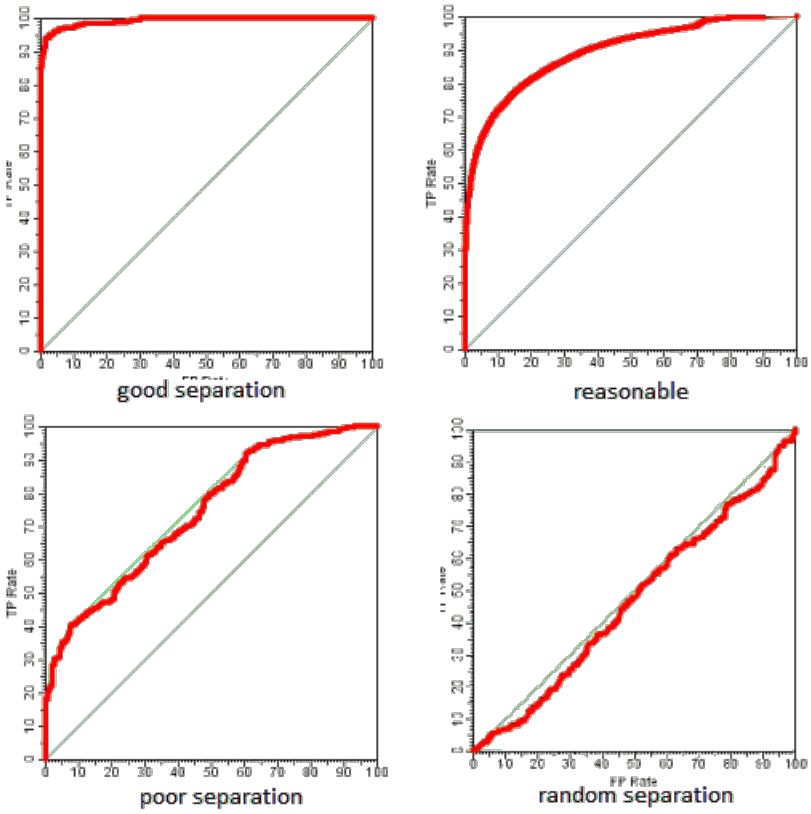

Figure 2: ROC кривые

Задача заключается в выборе точки, которая находится ближе всего к координате (0, 1). То есть цель уменьшить $FPR = \frac{FP}{FP+TN}$ и увеличить полноту(recall) $TPR = \frac{TP}{TP+FN}$. Выбор точки может осуществляться как поиском расстояния от каждой точки кривой до координаты (0, 1) или же точкой, через которую проходит касательная $y = \frac{neg}{pos}x$, где neg - доля классов с меткой False, и pos - доля классов с меткой True.

Оптимальная кривая, то есть кривая, соответствующая оптимальному классификатору, это кривая с наибольшей площадью(наибольший AUC).