1. изучите понятие ROC и ROC-AUC. Докажите, что площадь под ROC-кривой численно равна вероятности оценить объект положительного класса выше объекта отрицательного

Пусть X - результат предсказания (случайная величина), T - пороговый параметр (объект классифицируется как положительный, если X > T, и как отрицательный иначе). Если объект принадлежит положительному классу, то плотность вероятности X - $f_1(x)$, иначе $f_2(x)$. Тогда

чувствительность TPR = $\int_T^{\infty} f_1(x)dx$

FPR = $\int_T^{\infty} f_2(x)dx$

Найдём площадь под ROC-кривой, как интеграл от ф-ии, которая задана параметрически.

$$S = \int_{-\infty}^{+\infty} TPR(T)(FPR'(T))dT = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I(T' > T)f_1(T')f_2(T)dT'dT = P(X_1 > X_2)$$

где X_1 предсказание для объекта положительного класса, X_2 предсказание для объекте отрицательного класса.
Т.е. площадь под ROC-кривой численно равна вероятности оценить объект положительного класса выше объекта отрицательного.

2. приведите пример задачи, когда precision важней recall, когда recall важней precision и когда необходимо находить баланс между этими двумя величинами

Precision важнее, чем recall в задачах, где критичным является не ошибиться и не отнести объект из отрицательного класса к положительному, т.е. эта ошибка будет более критичной, чем отнесение положительного класса к отрицательному. Примером задачи может быть ,например, определение здоровым ли является донорский орган. В этой задаче важнее не пересадить больной донорский орган другому человеку, чем решить, что здоровый орган является больным и не пересаживать его никому. В этой ситуации точность важнее.

Recall важнее, чем precision, когда нам надо найти как можно больше объектов из положительного класса и то что отрицательный класс будет отнесен к положительному будет не так важно. Пример задачи: определение болен ли раком человек, в этой задаче надо найти как можно больше больных и если здоровые люди будут отнесены к больным это не так критично (т.к. можно будет провести еще дополнительные обследования) главное чтобы все больные были обнаружены.

Баланс между этими величинами надо искать когда нет жестких требований ни к точности, ни к полноте, но хотелось бы иметь хорошее решение. Например, определение тональности текста для рекомендации к прочтению этого документа для человека. В этой задаче не сильно страшно если ответ будет неправильный, так как целью всё равно является получить некоторую оценку текста, дать рекомендацию к

прочтению для человека, а дальше он уже сам может решить стоит ли читать документ. Но хотелось бы получить как можно более точное и полное решение, поэтому тут и необходимо найти баланс.