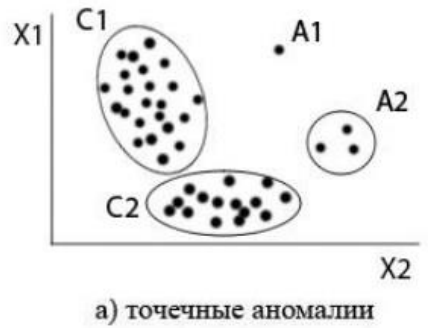
Введение в машинное обучение

Лекция 6: Оценка качества алгоритмов машинного обучения. Кросс-валидация. Поиск аномалий и артефактов в выборке.

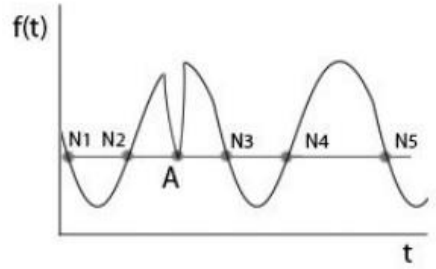
Полина Полунина

Поиск аномалий и артефактов в выборке

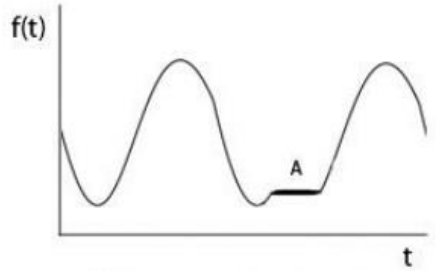
- Что такое аномалии и артефакты?
  - Пропуски в данных
  - Явно не верные/грязные значения, например возраст 689 лет или !*,»# лет
- Что с ними делать?
  - Удалять
  - Заполнять значениями
- Если заполнять значениями, то какими?
  - В случае с явно не верными значениями – по контексту, если это возможно
  - В случае с пропусками – mean, max, min, quantile…



а) точечные аномалии



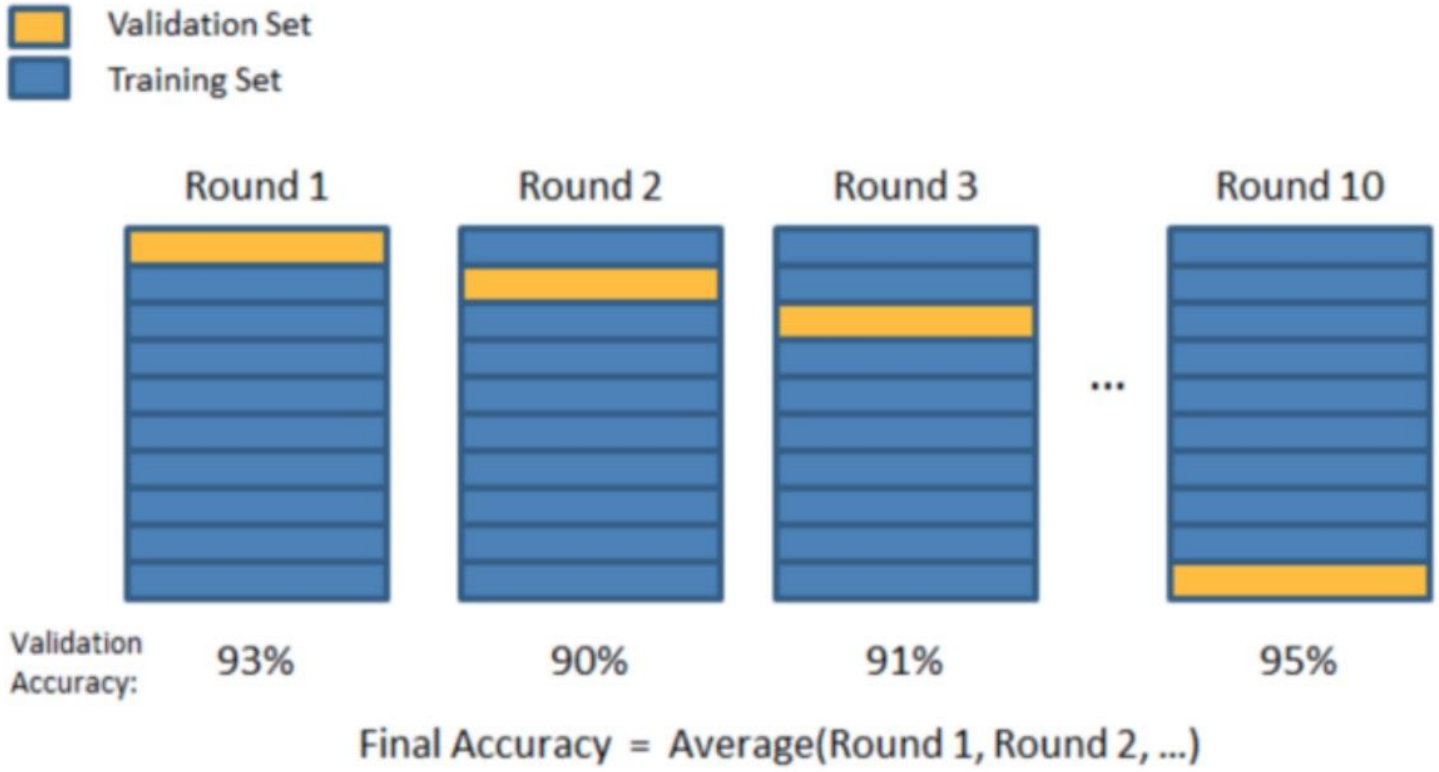
б) контекстуальные аномалии



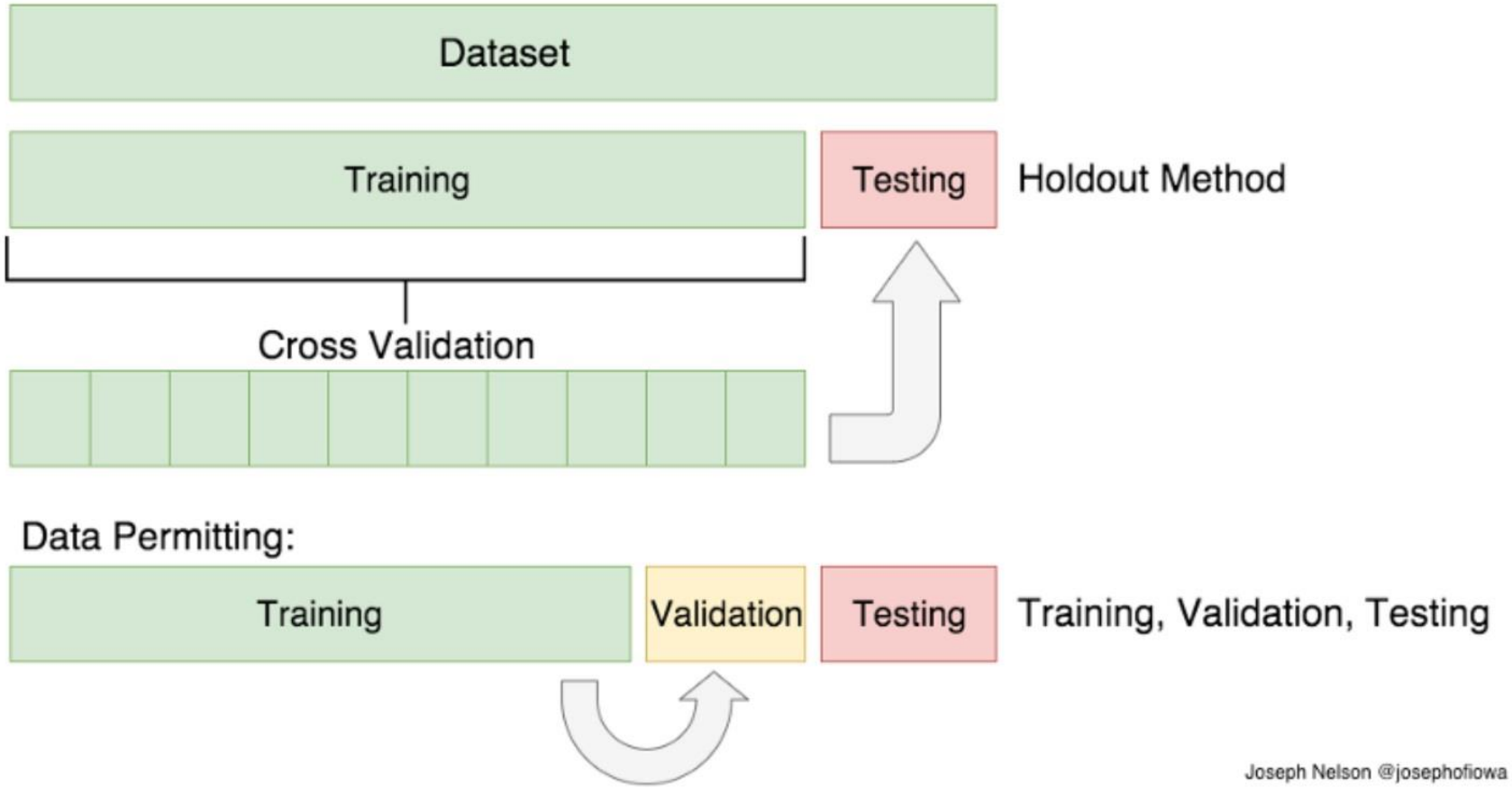
в) коллективные аномалии

Подход к разделению данных на подвыборки

- Основная идея: разделить выборку на несколько независимых частей, чтобы оценить обобщающую способность модели



Подход к разделению данных на подвыборки



K-Fold кросс-валидация: Варианты

- **Train/Test Split**: С одной стороны, k может быть равно 1, т.е. всего одно разделение на train/test
- **LOOCV**: С другой стороны, k может быть равно кол-ву наблюдений в датасете, т.е. предсказание делается каждый раз на одном наблюдении. Такой подход называется leave-one-out cross-validation
- **Stratified**: В задачах классификации с несбалансированными данными возникает потребность разбить данные на куски так, чтобы в каждом куске пропорция классов сохранялас
- **Repeated**: Разбиваем на k фолдов несколько раз

Оценка качества алгоритмов машинного обучения

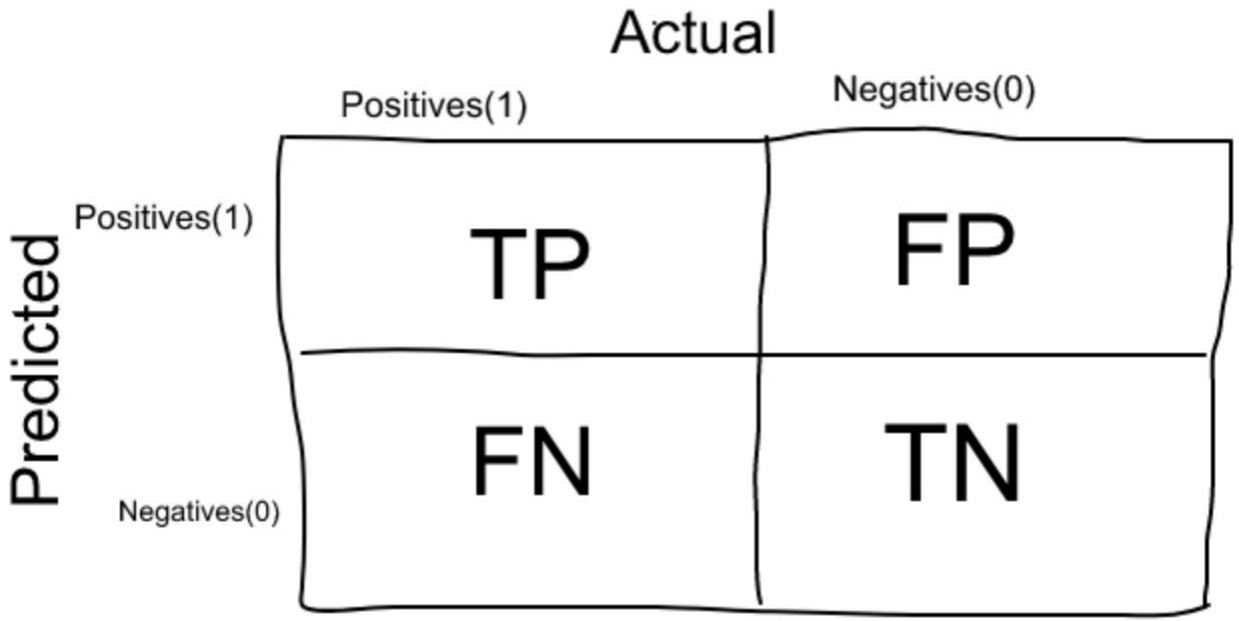
Задача Классификации

Confusion Matrix
TP Rate
TN Rate
FP Rate
FN Rate
Accuracy
Precision
Recall a.k.a. Sensitivity
Specificity
ROC_AUC
F1 Score
F-Beta Score
PR_AUC
Gini Coefficient
Log Loss…

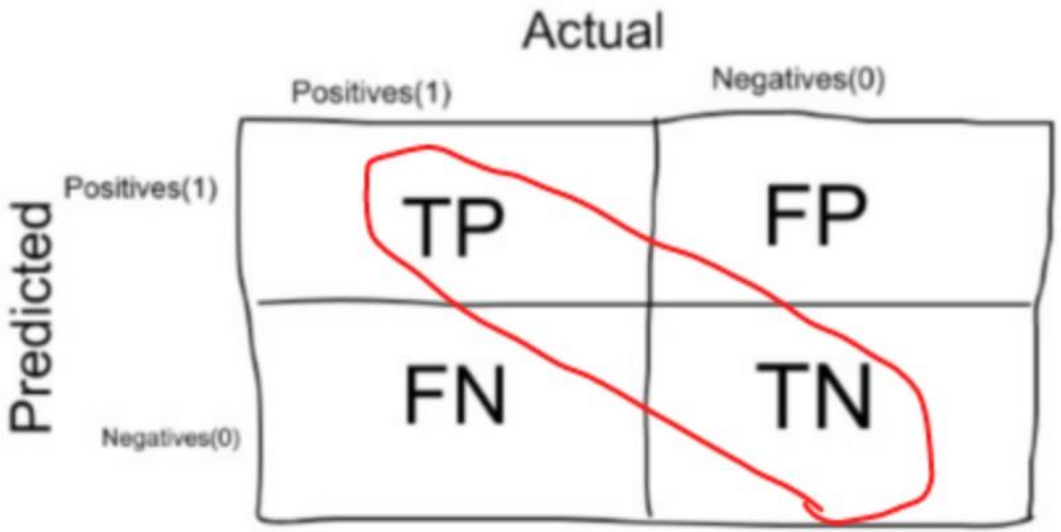
Задача Регрессии

R^2
R^2 adjusted
MAE
MAPE
MSE
RMSE

Метрики для задачи классификации: Confusion Matrix

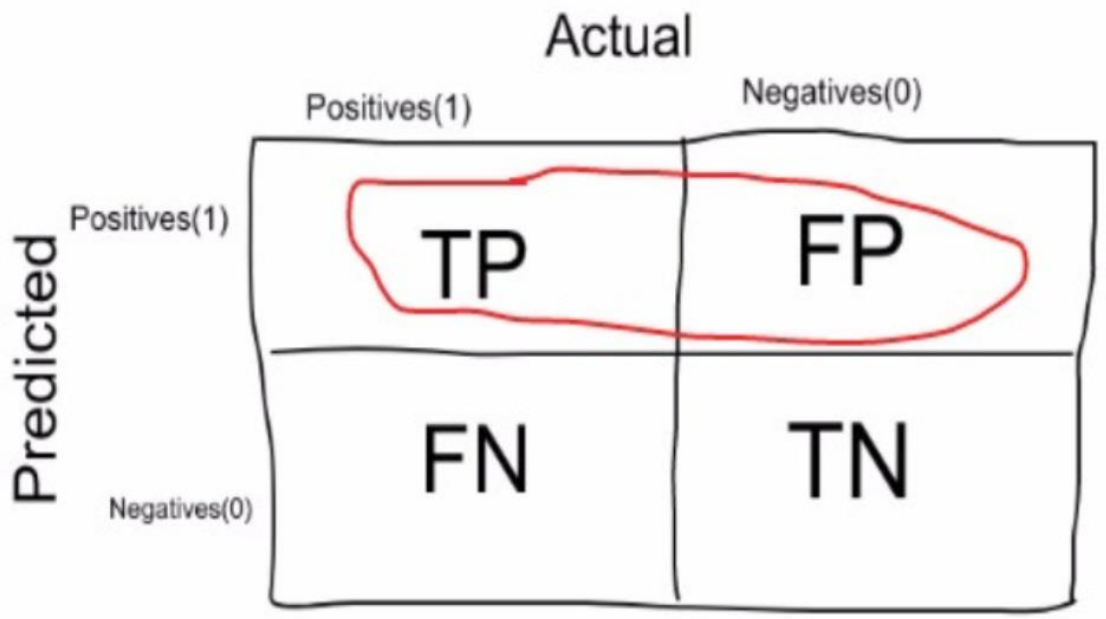


Метрики для задачи классификации: Confusion Matrix



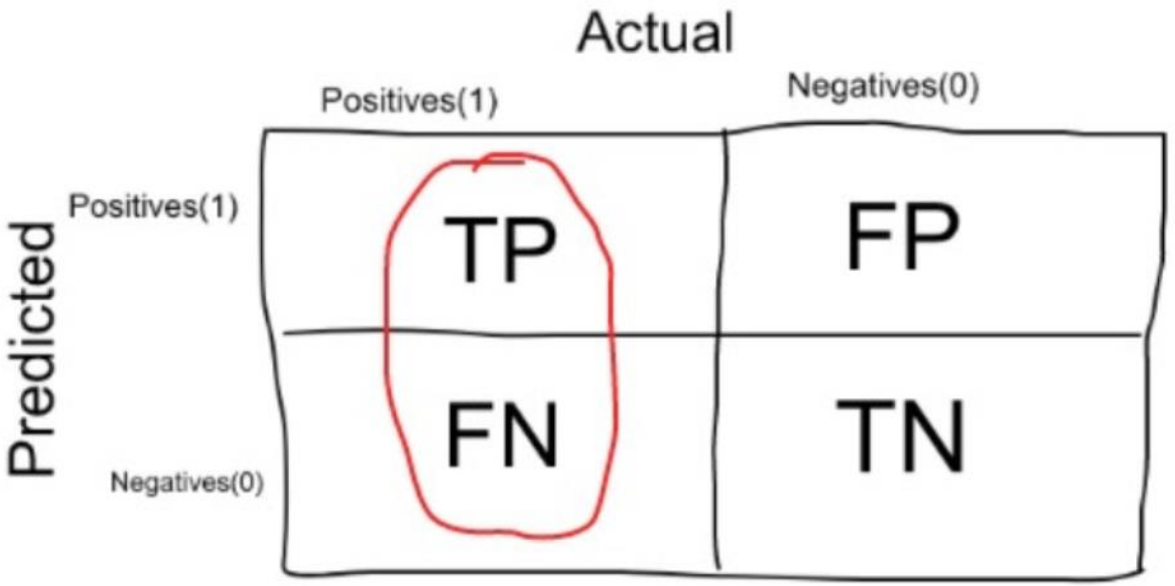
$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Метрики для задачи классификации: Confusion Matrix



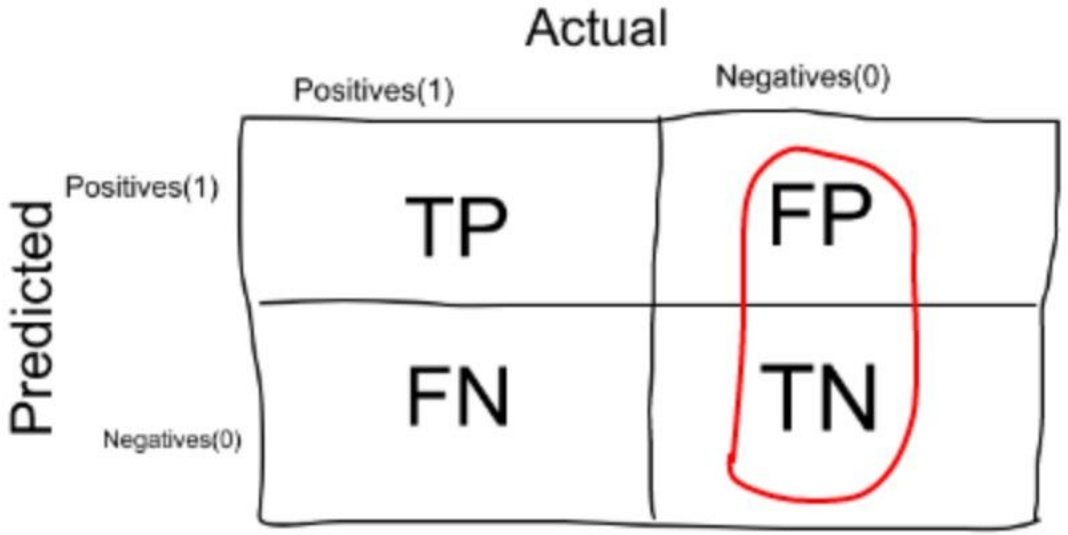
$$\text{Precision} = \frac{TP}{TP + FP}$$

Метрики для задачи классификации: Confusion Matrix



$$\text{Recall} = \frac{TP}{TP + FN}$$

Метрики для задачи классификации: Confusion Matrix

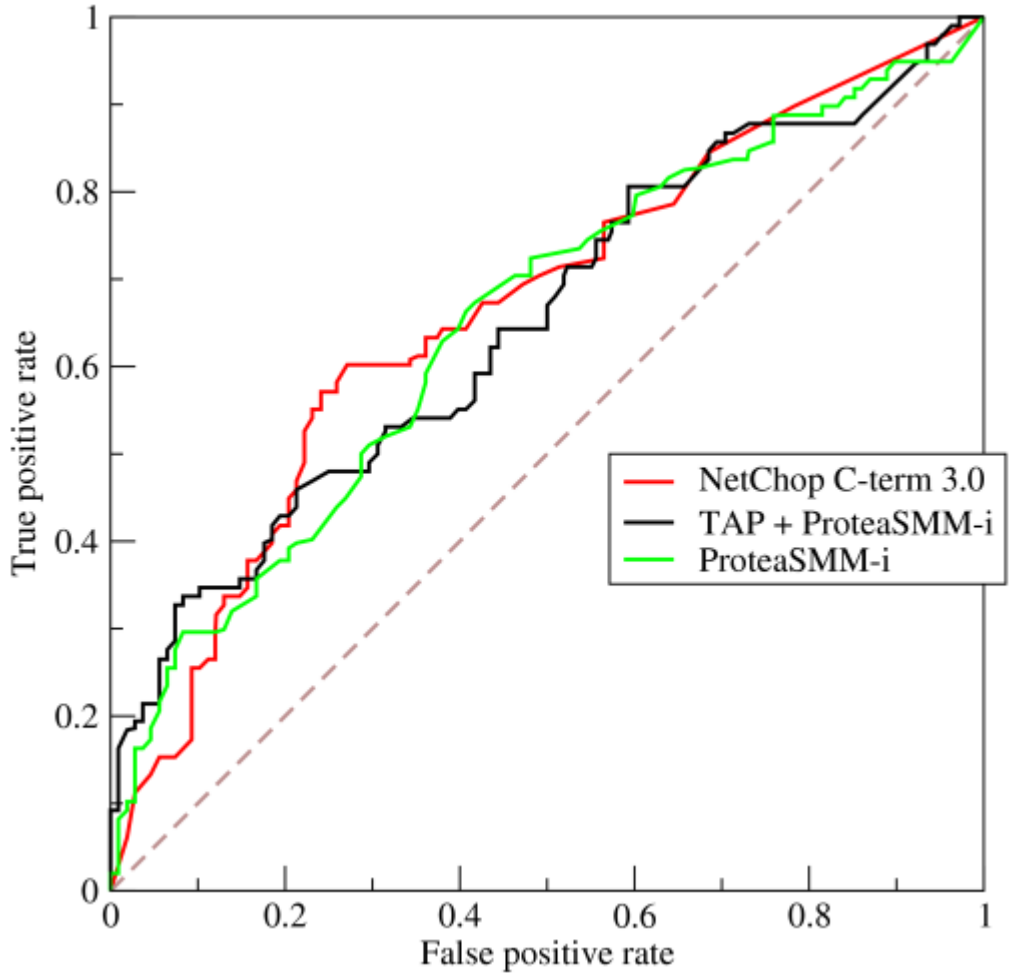


$$\text{Specificity} = \frac{TN}{TN + FP}$$

Метрики для задачи классификации:

- F1 Score = 2 * Precision * Recall / (Precision + Recall)
- $$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}$$
- ROC_AUC - площадь под ROC кривой
- Gini = 2 * AUC - 1
- PR_AUC – площадь под Precision-Recall кривод

Метрики для задачи классификации: ROC кривая



Метрики для задачи регрессии:

- $\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$
- $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$
- RMSE = sqrt(RMSE)
- MAPE = $\frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$