

# Машинное обучение и майнинг данных

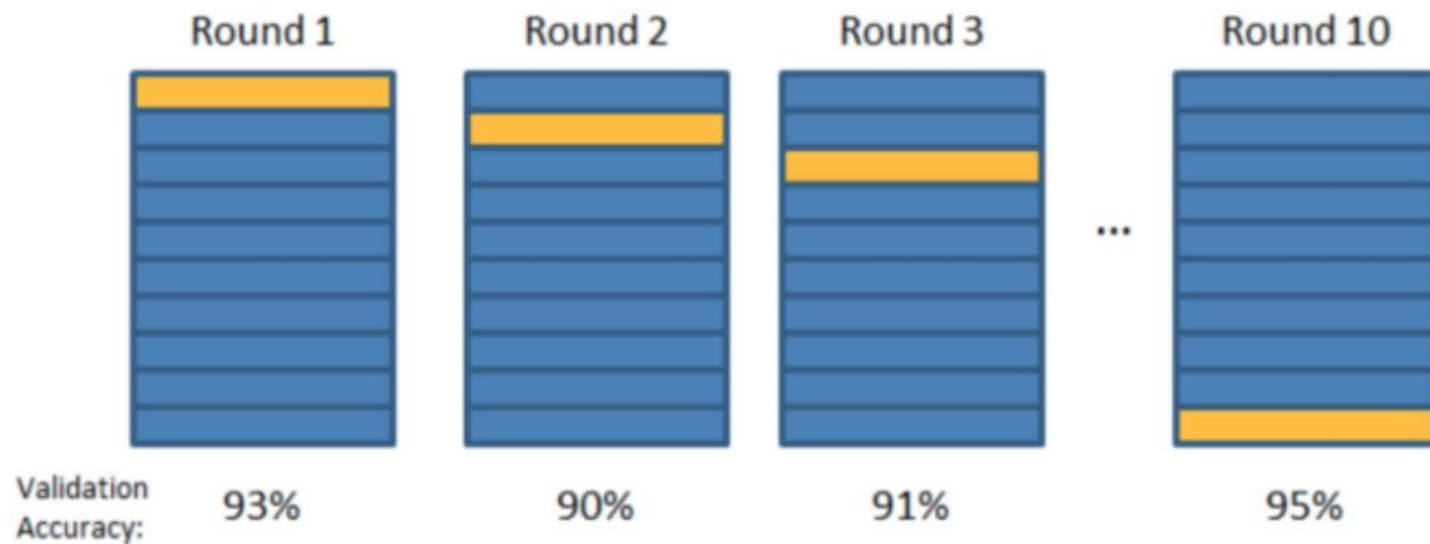
Лекция 6: Оценка качества алгоритмов машинного обучения.  
Кросс-валидация.

Полина Полунина

# Подход к разделению данных на подвыборки

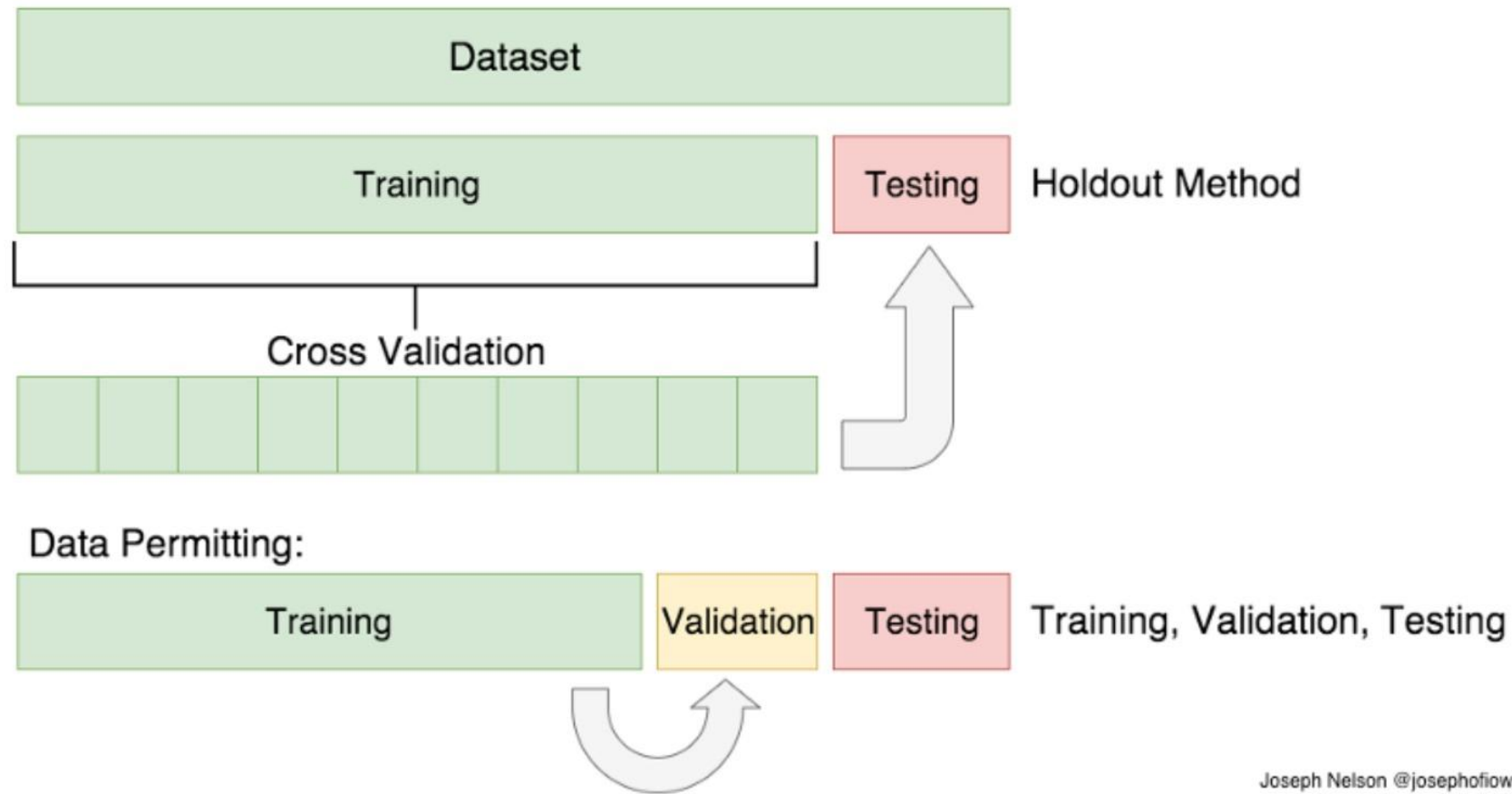
- Основная идея: разделить выборку на несколько независимых частей, чтобы оценить обобщающую

СПС



Final Accuracy = Average(Round 1, Round 2, ...)

# 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 сохранялась
- **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

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

# Метрики для задачи классификации:

## Confusion Matrix

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

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

# Метрики для задачи классификации:

## Confusion Matrix

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

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



# Метрики для задачи классификации:

## Confusion Matrix

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

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

# Метрики для задачи классификации:

## Confusion Matrix

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

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

# Метрики для задачи классификации:

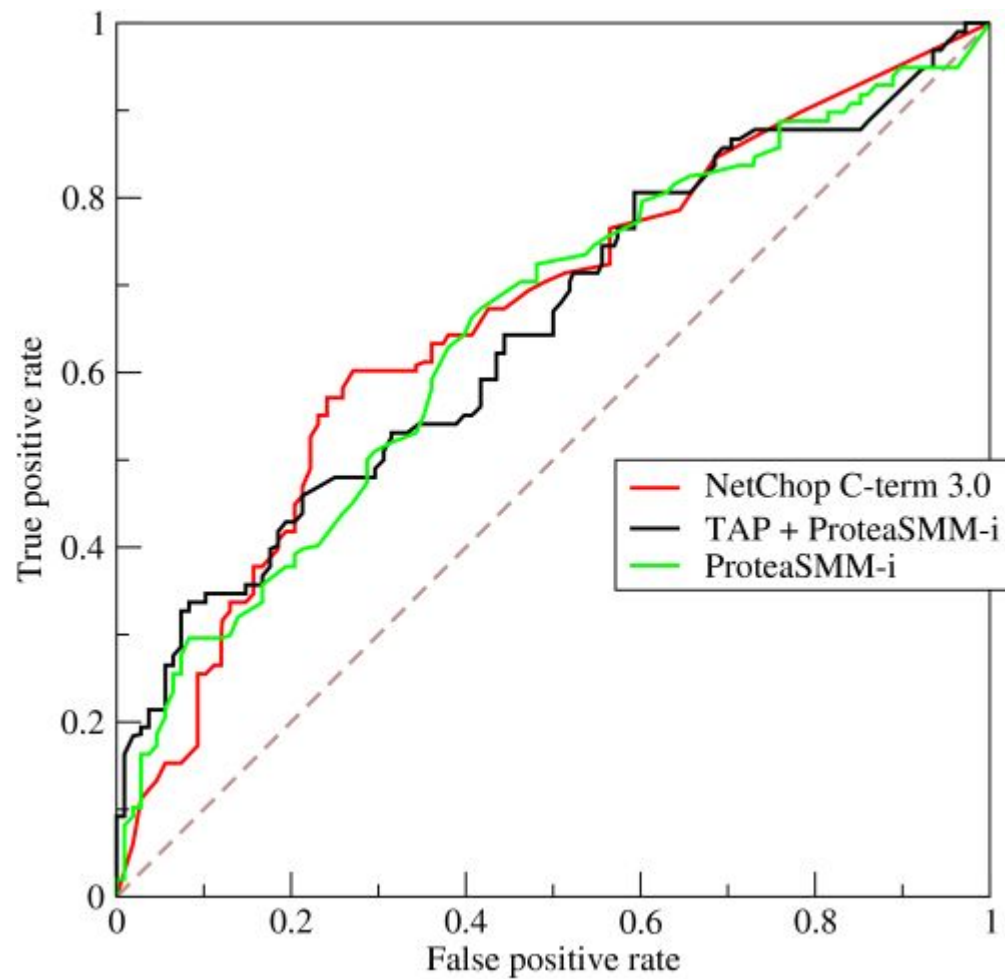
- F1 Score =  $2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$

$$F_{\beta} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

- ROC\_AUC - площадь под ROC кривой
- Gini =  $2 * \text{AUC} - 1$
- PR AUC – площадь под Precision-Recall кривод

# Метрики для задачи классификации:

## ROC кривая



# Метрики для задачи регрессии:

- $$\text{MAE} = \frac{\sum_{i=1}^n |y_i - x_i|}{n} = \frac{\sum_{i=1}^n |e_i|}{n}$$

- $$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

- $$\text{RMSE} = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$$

- $$\text{MAPE} =$$