Данные

Типы:

1) Изображения
2) Видео
3) Табличные
4) Текстовые
5) Звуковые

Источники:

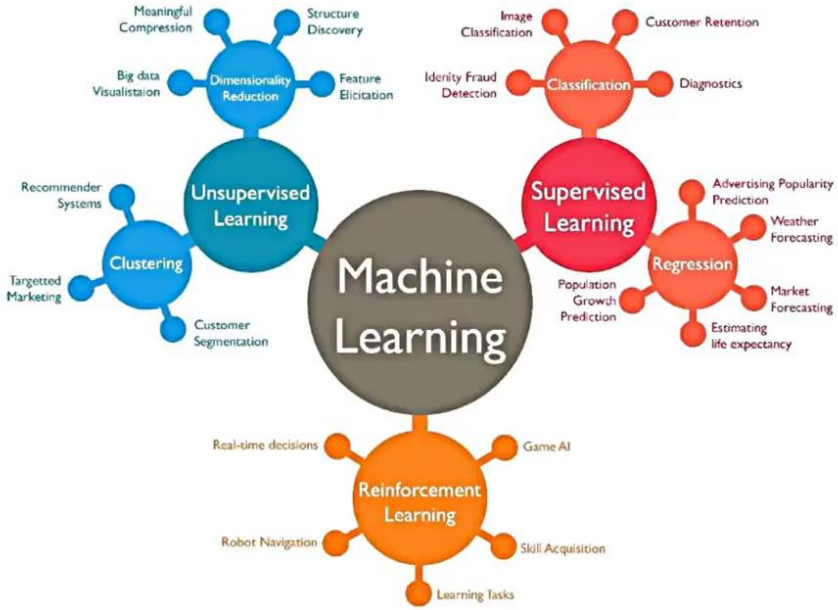
1) Наборы данных от заказчика
2) Медицинские организации
3) Государственные структуры
4) Банки
5) Научные исследования

Машинное обучение

Машинное обучение (Machine Learning) - область искусственного интеллекта. Data-driven подход, полагающийся на поиск закономерностей в имеющихся размеченных или неразмеченных данных. Идея в том, чтобы не программировать алгоритм решения задачи вручную, а “выучить” его из данных.

Машинное обучение разделяется на несколько основных подходов:

- Обучение с учителем (supervised learning)
  - Классификация (classification)
  - Регрессия (regression)
  - Ранжирование (learning to rank)
- Обучение без учителя (unsupervised learning)
  - Кластеризация (clustering)
  - Уменьшение размерности (dimensionality reduction)
- Обучение с частичным привлечением учителя (semi-supervised learning)
- Обучение с подкреплением (reinforcement learning)



Метрики качества

Метрики качества для классификации:

1) Accuracy
2) Confussion matrix, recall, precision, F1
3) ROC AUC

Метрики качества для регрессии:

1) Mean Absolute Error (MAE)
2) Mean Squared Error (MSE)
3) Root Mean Squared Error (RMSE)

Метрики качества для ранжирования:

1) Kendall’s tau

Тренировочные, валидационная, тестовые данные

Train – данные для обучения,
Valid – данные для проверки модели при обучении,
Test – данные для проверки модели после обучения.

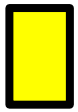
Исходный набор данных делится на Train/Valid в соотношении 90%/10% или 80%/20%.

Валидация дает понять насколько хорошо обучена модель и нет ли переобучения

Использовать одни и те же данные для обучения и тестирования недопустимо, т.к. модель должна знать, как будет работать метод на новых данных.

Перекрёстная проверка (cross validation)

– train + valid



– test

Можно 100% выборки разделить на 4 части (блока) и одну из них (25 %) выделить под тестирование. Какой именно из блоков передать на тестирование?

Метод ML

Верно 5
Неверно 1

Метод ML

Верно 4
Неверно 2

Метод ML

Верно 1
Неверно 5

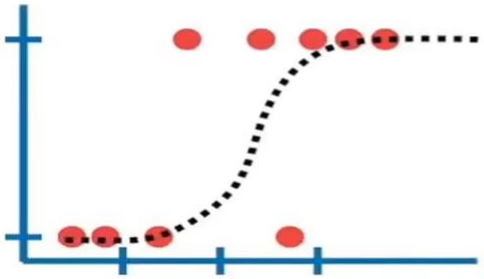
Метод ML

Верно 6
Неверно 0

Перекрёстная проверка (cross validation)

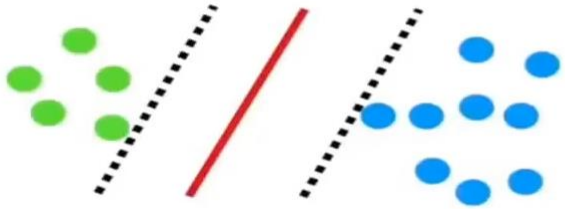
Проводим тестирование для каждого метода ML. Результаты всех 4 тестов складываем и выбираем лучший метод:

Логистическая регрессия



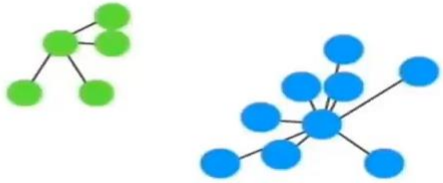
Верно 16
Неверно 8

Метод опорных векторов (SVM)



Верно 18
Неверно 6

Метод K-ближайших соседей



Верно 10
Неверно 12

Была произведена **Четырехкратная перекрестная проверка** (**4-Fold Cross Validation**). В общем случае число блоков произвольное.
Также возможна **поэлементная перекрестная проверка** (**Leave One Out Cross Validation**).

Confusion matrix

Пусть есть 2 класса объектов $y\_true \in \{0,1\}$ и их предсказанные значения $y\_pred \in \{0,1\}$

Confusion matrix (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html):

True Negative (TN)	False Positive (FP)
False Negative (FN)	True Positive (TP)

True Negative — Классификатор верно классифицировал 0 как 0
False Positive — Классификатор ложно классифицировал 0 как 1
False Negative — Классификатор ложно классифицировал 1 как 0
True Positive — Классификатор верно классифицировал 1 как 1

True Negative == предсказан 0 (Negative), и предсказан верно (True)

```
In [15]: y_true = [0, 0, 0, 1, 1, 1, 1, 1]
         y_pred = [0, 1, 0, 1, 0, 1, 0, 1]
         confusion_matrix(y_true, y_pred)

Out[15]: array([[2, 1],
                [2, 3]], dtype=int64)
```

		y_pred (j)	
		0	1
y_true (i)	0	2	1
	1	2	3

$C_{i,j}$ – количество объектов класса i, которым предсказан класс j

Confusion matrix

3 класса объектов $y\_true \in \{0,1,2\}$ и их предсказанные значения $y\_pred \in \{0,1,2\}$

```
>>> from sklearn.metrics import confusion_matrix
>>> y_true = [2, 0, 2, 2, 0, 1]
>>> y_pred = [0, 0, 2, 2, 0, 2]
>>> confusion_matrix(y_true, y_pred)
array([[2, 0, 0],
       [0, 0, 1],
       [1, 0, 2]])
```

		y_pred (j)		
		0	1	2
y_true (i)	0	2	0	0
	1	0	0	1
	2	1	0	2

$C_{i,j}$ – количество объектов класса *i*, которым предсказан класс *j*

Метрики

Пусть есть 2 класса объектов $y\_true \in \{0,1\}$ и их предсказанные значения $y\_pred \in \{0,1\}$

True Negative (TN)	False Positive (FP)
False Negative (FN)	True Positive (TP)

Accuracy – доля правильных ответов

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision (точность) – отношение количества объектов целевого класса, которые классифицированы как объекты этого класса, к общему количеству объектов, которые классифицированы как объекты этого класса.

$$precision = \frac{TP}{TP + FP}$$

Recall (полнота) – отношение количества объектов, которые классифицированы как объекты целевого класса, к общему количеству объектов этого класса.

$$recall = \frac{TP}{TP + FN}$$

Среднее гармоническое *f1_score*

$$f1\_score = \frac{2*(precission*recall)}{precission+recall}$$

В https://habr.com/company/ods/blog/328372/ и https://en.wikipedia.org/wiki/Confusion_matrix confusion matrix составлена с другим порядком элементов.

metrics

```
from sklearn import metrics

y_pred = model.predict_classes(X_test_nn, batch_size=1)
mtrs = metrics.classification_report(y_test_oh_n, y_pred)
print(mtrs)
```

	precision	recall	f1-score	support
0	0.99	1.00	1.00	160
1	1.00	1.00	1.00	160
2	1.00	0.99	1.00	160
avg / total	1.00	1.00	1.00	480

Метрики: https://habr.com/company/ods/blog/328372/