

Лабораторная работа № 4

Тема: Линейные классификаторы.

В качестве данных для обучения можно взять свой датасет из предыдущей лабораторной работы или любой интересный датасет, подходящий для бинарной классификации.

1. Обучите на своих данных **модель логистической регрессии LogisticRegression ()**.

- Рассчитайте точность на обучающих и тестовых данных:

```
print("Правильность на обучающем наборе: {:.2f}".format(model.score(X_train, y_train)))  
print("Правильность на тестовом наборе: {:.2f}".format(model.score(X_test, y_test)))  
где model – это ваша обученная модель
```

- Измените в модели параметр регуляризации C, по умолчанию C=1, пересчитайте с C= 100 и C= 0,01. Вновь выведите правильность на обучающем и тестовом наборах. Сделайте выводы.
- Добавьте в модель L2-регуляризацию, по умолчанию используется L1-регуляризация:

```
LogisticRegression(penalty='l2', C=0.1)
```

- Рассчитайте метрики качества (*accuracy, precision, recall*) и матрицу ошибок для наилучшей на ваш взгляд модели.

2. Обучите на своих данных **модель метода опорных векторов SVC()**.

- Рассчитайте точность на обучающих и тестовых данных.
- В методе опорных векторов для регуляризации в основном используют параметры C и гамма. При помощи метода GridSearchCV найдите наилучшую комбинацию этих параметров:

```
from sklearn.model_selection import GridSearchCV  
SVC_params = {"C": [0.1, 1, 10], "gamma": [0.2, 0.6, 1]}  
SVC_grid = GridSearchCV(model_SVC, SVC_params, cv=5, n_jobs=-1)  
SVC_grid.best_score_, SVC_grid.best_params_
```

- Выведите точность этой модели на обучающих и тестовых данных, сделайте выводы
- Рассчитайте метрики качества (*accuracy, precision, recall*) и матрицу ошибок для наилучшей модели метода опорных векторов.

3. В этом же файле (блокноте) обучите на этом же датасете модели дерева решений и K-ближайших соседей. Выведите их точность на обучающих и тестовых данных.

4. Постройте в одних осях четыре ROC-кривые для 4-х обученных моделей (логистической регрессии, метода опорных векторов, дерева решений и K-ближайших соседей). Сделайте вывод какая модель лучше делает прогноз на ваших данных.

```
from sklearn.metrics import RocCurveDisplay  
ax = plt.gca()  
rfc_disp = RocCurveDisplay.from_estimator(model_SVC, X_test, y_test, ax=ax)
```

```
svc_disp = RocCurveDisplay.from_estimator(model_DT, X_test, y_test, ax=ax)
... ..
... ..
```

```
plt.show()
```

где `model_` - это ваши ранее обученные 4 модели

Вопросы

1. В чем суть линейных классификаторов? Какие методы относятся к линейным классификаторам?
2. Что такое регуляризация?
3. Объясните назначение L1 и L2-регуляризации?
4. Объясните влияние параметров C и гамма на регуляризацию модели?
5. Какой метод помогает подобрать лучшую комбинацию параметров?
6. Что такое специфичность и чувствительность модели?
7. Как по ROC-кривой можно сравнить модели. Что означает AUC в ROC-кривой?