Лабораторная 3: Линейные методы

1. Классификация

Метрики

Все описанные ниже метрики реализованы в модуле `sklearn.metrics`.

Рассмотрим задачу двухклассовой классификации. Пусть $f_w(x) = \langle w, x \rangle$ — некоторая линейная модель, для которой правило классификации по порогу w_0 записывается в виде $F_{w,w_0}(x) = \text{sign}\,(f_w(x) - w_0)$.

Пусть $X_{test} = (X_i, Y_i)_{i=1,\ldots,n}$ — тестовая выборка, причем $Y_i \in \{-1, +1\}$. Обозначим $I_{good} = \{i | Y_i = 1\}$, $I_{bad} = \{i | Y_i = -1\}$ — индексы хороших и плохих объектов. Для данного классификатора F обозначим так же $I^F_{good} = \{i | F(X_i) = 1\}$, $I^F_{bad} = \{i | F(X_i) = -1\}$ — индексы объектов, которые классификатором F классифицируются как хорошие и плохие.

В данной модели можно рассмотреть следующие метрики качества.

1. Precision (Точность) — доля действительно хороших объектов среди классифицируемых как хорошие

$$Prec(F, X_{test}) = \frac{\left|I_{good} \cap I^F_{good}\right|}{\left|I^F_{good}\right|}.$$

2. Recall (Полнота) — доля объектов, классифицируемых как хорошие, среди действительно хороших объектов

$$Recall(F, X_{test}) = \frac{\left|I_{good} \cap I^F_{good}\right|}{|I_{good}|}.$$

3. F_1**.** Ясно, что в некотором смысле две предыдущие метрики противоречат друг другу. Например, если все документы классифицировать как хорошие. Поэтому разумно эти метрики смешать

$$F_1 = \frac{2}{\frac{1}{Prec} + \frac{1}{Recall}}.$$

4. ROC-AUC Доля ложных положительных классификаций (False Positive Rate, FPR):

$$\text{FPR}(F, X_{test}) = \frac{\sum_{i=1}^n I\{F(X_i) = 1, Y_i = -1\}}{\sum_{i=1}^n I\{Y_i = -1\}}$$

Доля верных положительных классификаций (True Positive Rate, TPR):

$$\mathrm{TPR}(F, X_{test}) = \frac{\sum_{i=1}^{n} I\{F(X_i) = 1, Y_i = 1\}}{\sum_{i=1}^{n} I\{Y_i = 1\}}.$$

ROC-кривой называется график зависимости TPR от FPR при изменении параметра w_0, и проходит через точки (0, 0) и (1, 1). Чем выше лежит кривая, тем лучше качество классификации. Метрика AUC (Area Under Curve) есть площадь под ROC-кривой.

5. MSE Подходит для задачи регрессии, но может давать адекватные значения и для задач классификации.

$$MSE(F, X_{test}) = \frac{1}{|X_{test}|} \sum_{(x,y) \in X_{test}} (F(x) - y)^2$$

В отличии от всех предыдущих метрик она является не мерой качества, а мерой ошибки.

Задание

Цель лабораторной — исследовать, как различные модели влияют на качество в зависимости от выбранной метрики.

Рассмотрим следующие модели классификации

1. LDA — линейный дискриминантный анализ (принцип максимума апостериорной вероятности, в котором компоненты имеют гауссовское распределение с одинаковой матрицей ковариаций). В старых версиях это `sklearn.lda.LDA`, в новых `sklearn.discriminant_analysis.LinearDiscriminantAnalysis`.
2. Логистическая регрессия (`sklearn.linear_model.LogisticRegression`)
3. SVM (`sklearn.svm.LinearSVC`, либо `sklearn.svm.SVC` с параметром `kernel="linear"`)

Каждую модель классификации можно записать в виде $F_{w,w_0}(x) = \mathrm{sign}\left(\langle w, x\rangle - w_0\right)$, проверьте это для каждой из них.

Сгенерируйте двумерную выборку размера 500 из двух классов, для этого можно воспользоваться `sklearn.datasets.make_blobs`. Проверьте, что классы достаточно хорошо пересекаются, но не слишком сильно. Этого можно добиться, например, изменением параметра `cluster_std`, который отвечает за дисперсию кластеров. Разбейте выборку поровну на трейн и тест (пользуйтесь `sklearn.cross_validation.train_test_split`)

1. (3 балла) Обучите на трейне предложенные выше модели, получив в каждой из них параметры w и w_0. Для вектора w у моделей есть поле `coef_`, а для числа w_0 поле `intercept_`. Для SVM здесь лучше воспользоваться `sklearn.svm.LinearSVC`. Для каждой модели постройте графики метрик (Prec, Recall, F_1, MSE) в зависимости от w_0 при фиксированном w. Для каждой модели должен быть свой график, на котором вместе изображены зависимости для всех метрик. Стоит заметить, что для построения графика не нужно использовать сетку из значений w_0, поскольку при монотонном увеличении w_0 классификация меняется не более n раз, где n — размер тестовой выборки. Поскольку метрики дискретны (в данном случае даже MSE :)), то и график должен выглядеть как кусочно-постоянная функция, а не кусочно-линейная, учтите это при построении графика. После построения всех графиков не забудте сделать выводы. Выводы должны содержать ответы на вопросы: как ведут себя метрики по отношению друг к другу, какая модель оказалась лучшей, почему MSE в данном случае ведет себя адекватно, да и вообще, почему она тут дискретна ...

2. (1 балл) Проведите аналогичное исследования для случая, когда классы хорошо разделяются, и для случая, когда классы сильно перемешаны.

3. (2 балла) Теперь для каждой модели посчитайте метрику ROC-AUC и постройте график ROC-кривой (`sklearn.metrics.roc_curve`). При построении графика учтите, что в функцию нужно передавать не сами предсказания, а вероятности для первого класса. Оценку вероятностей можно получить с помощью функции `predict_proba`. В случае SVM это возможно только для `sklearn.svm.SVC`, если указать `probability=True`. Сделайте выводы.

4. (2 балла) Рассмотрим теперь SVM. Для некоторых значений C обучите модель на обучающей выборке и визуализируйте полученную классификацию так, как показано в ноутбуке с семинара 11. Постройте графики метрик F_1 и ROC-AUC в зависимости от C на тестовой выборке. Сделайте выводы.

2. Регрессия

Рассматриваем только метрику MSE. Скачайте данные с Диска и разбейте из на трейн и тест в сооотношении 3:1. Реализуйте самостоятельно (в одну строчку) и обучите линейную регрессию. Посчитайте ошибку MSE на тесте. Проверьте значение детерминанта матрицы, которую приходится обращать в процессе обучения. Надежный ли получился результат?

5. (2 балла) Обучите Ridge-регрессию для различных значений параметра регуляризации α. Постройте график метрики MSE в зависимости от значения α. Проведите аналогичный эксперимент для LASSO-регрессии и ElasticNet-регрессии. В последнем случае параметр изменяйте так, как описано в ноутбуке с семинара 11. Сделайте выводы.

6. (1 балл) Для Ridge-регрессии и LASSO-регрессии постройте графики траекторий путей весов признаков так, как показано в ноутбуке с семинара 11. Сделайте выводы.

7. (3 балла) SVM-регрессия (`sklearn.svm.SVR`) Какую функцию потерь использует SVM-регрессия? Постройте графики метрик MSE и среднего значения функции потерь SVM-регрессии на обучающей и тестовой выборках в зависимости от C для фиксированного ε и в зависимости от ε при фиксированном C. Подберите оптимальные значения параметров ε и C. Сделайте выводы.