Задания на курсовую работу
по предмету «Методы машинного обучения»

Программы курсовой работы должны быть написаны на языке Python с использованием библиотек sklearn, pandas, numpy, matplotlib (допускается использование других библиотек).

Теоретический материал для выполнения курсовой работы можно посмотреть на сайте в текстовом варианте:

https://proproprogs.ru/ml

или в формате видео лекций:

http://tk.ulstu.ru/video.php?id=3#video

Порядок выполнения работы

1. Скачать и изучить структуру датасета, представленного в виде текстового файла в формате csv (адрес и название файла указано в списке вариантов), для задачи бинарной классификации образов.

2. Выполнить загрузку csv файла в программе на Python, используя библиотеку pandas. Удалить из данных неинформативные столбцы (например, порядковый номер, дату формирования записи и т.п.).

3. На основе загруженных данных сформировать наборы векторов x_train, y_train обучающей выборки (в виде массивов numpy) и x_test, y_test – тестовой (отложенной) выборки. Разбиение образов из датасета делать случайным образом в пропорции 70% на 30% (70% - обучающая выборка; 30% - тестовая выборка).

4. Если обучающая выборка не сбалансирована (образов одного класса заметно больше или меньше другого класса), то сбалансировать ее путем переноса части образов класса большей мощности в тестовую выборку.

5. Для линейной модели бинарной классификации:

$$a(x) = sign(\langle \omega, x \rangle) = sign\left(\sum_{i=1}^{n} \omega_i x_i + \omega_0\right)$$

найти вектор весовых коэффициентов $\omega = [\omega_0, \omega_1, \omega_2, ..., \omega_n]^T$ с помощью алгоритма стохастического градиентного спуска (SGD) по обучающей выборке (алгоритм SGD реализовывать самостоятельно, не брать готовую реализацию из библиотек). Функцию потерь брать в соответствии со своим вариантом. Гиперпараметры алгоритма стохастического градиентного спуска

(шаг сходимости, максимальное число итераций и др.) выберите самостоятельно так, чтобы получить наименьшее значение среднего эмпирического риска $Q\left(X^l\right)$.

6. Выполнить тестирование полученного алгоритма $a(x)$ на тестовых данных. Вычислить метрику accuracy, средний эмпирический риск по полученным результатам.

7. Добавить L2-регуляризатор в алгоритм стохастического градиентного спуска и повторить процесс обучения и тестирования. Сравнить результаты с предыдущим случаем.

8. Оценить степень информативности признаков обучающей выборки путем вычисления собственных значений корреляционной матрицы признаков (см. метод главных компонент).

9. Обучить алгоритм SVM (метод опорных векторов) с линейным ядром по обучающей выборке и выполнить его тестирование по отложенной выборке. При этом вычислить метрику accuracy и средний эмпирический риск.

10. Обучить алгоритм SVM (метод опорных векторов) с полиномиальным ядром (степени 2) по обучающей выборке и выполнить его тестирование по отложенной выборке. При этом вычислить метрику accuracy и средний эмпирический риск.

Варианты заданий

№	Датасет	Функция потерь	Производная функции потерь
1	Airline Passenger Satisfaction.zip (целевой параметр: satisfaction)	$L(M)=\log_2\left(1+e^{-M}\right)$	$\frac{\partial L(M)}{\partial \omega}=-\frac{e^{-M}\cdot x^T\cdot y}{\left(1+e^{-M}\right)\cdot \ln 2}$
2	Bank Marketing Dataset.zip (целевой параметр: deposit)	$Q(M)=(1-M)^2$	$\frac{\partial Q(M)}{\partial \omega}=-2\cdot\left(1-\omega^T\cdot x\cdot y\right)\cdot x^T\cdot y$
3	Banking Dataset - Marketing Targets.zip (целевой параметр: deposit)	$S(M)=2\cdot\left(1+e^{M}\right)^{-1}$	$\frac{\partial S(M)}{\partial \omega}=-\frac{2\cdot e^{M}\cdot x^T\cdot y}{\left(1+e^{M}\right)^2}$
4	Brain Tumor.zip (целевой параметр: Class)	$E(M)=e^{-M}$	$\frac{\partial E(M)}{\partial \omega}=-e^{-M}\cdot x^T\cdot y$

5	Breast Cancer Prediction Dataset.zip (целевой параметр: diagnosis)	$L(M)=\log_2\left(1+e^{-M}\right)$	$\frac{\partial L(M)}{\partial \omega}=-\frac{e^{-M}\cdot x^T\cdot y}{\left(1+e^{-M}\right)\cdot \ln 2}$
6	Company Bankruptcy Prediction.zip (целевой параметр: Bankrupt)	$Q(M)=(1-M)^2$	$\frac{\partial Q(M)}{\partial \omega}=-2\cdot\left(1-\omega^T\cdot x\cdot y\right)\cdot x^T\cdot y$
7	Credit Card Fraud Detection.zip (целевой параметр: Class)	$S(M)=2\cdot\left(1+e^{M}\right)^{-1}$	$\frac{\partial S(M)}{\partial \omega}=-\frac{2\cdot e^{M}\cdot x^T\cdot y}{\left(1+e^{M}\right)^2}$
8	Diabetes Health Indicators Dataset.zip (целевой параметр: Diabetes_binary)	$E(M)=e^{-M}$	$\frac{\partial E(M)}{\partial \omega}=-e^{-M}\cdot x^T\cdot y$
9	E-Commerce Shipping Data.zip (целевой параметр: Reached.on.Time_Y.N)	$L(M)=\log_2\left(1+e^{-M}\right)$	$\frac{\partial L(M)}{\partial \omega}=-\frac{e^{-M}\cdot x^T\cdot y}{\left(1+e^{-M}\right)\cdot \ln 2}$
10	NASA - Nearest Earth Objects.zip (целевой параметр: hazardous)	$Q(M)=(1-M)^2$	$\frac{\partial Q(M)}{\partial \omega}=-2\cdot\left(1-\omega^T\cdot x\cdot y\right)\cdot x^T\cdot y$
11	Personal Key Indicators of Heart.zip (целевой параметр: HeartDisease)	$S(M)=2\cdot\left(1+e^{M}\right)^{-1}$	$\frac{\partial S(M)}{\partial \omega}=-\frac{2\cdot e^{M}\cdot x^T\cdot y}{\left(1+e^{M}\right)^2}$
12	Smoke Detection Dataset.zip (целевой параметр: Fire Alarm)	$E(M)=e^{-M}$	$\frac{\partial E(M)}{\partial \omega}=-e^{-M}\cdot x^T\cdot y$
13	Stellar Classification Dataset.zip (целевой параметр: class)	$L(M)=\log_2\left(1+e^{-M}\right)$	$\frac{\partial L(M)}{\partial \omega}=-\frac{e^{-M}\cdot x^T\cdot y}{\left(1+e^{-M}\right)\cdot \ln 2}$
14	Water Quality.zip (целевой параметр: Potability)	$Q(M)=(1-M)^2$	$\frac{\partial Q(M)}{\partial \omega}=-2\cdot\left(1-\omega^T\cdot x\cdot y\right)\cdot x^T\cdot y$

P.S. Здесь $M=\omega^T\cdot x\cdot y=y\cdot\left(\sum_{i=1}^{n}\omega_i\cdot x_i+\omega_0\right),\ y\in\{-1,+1\}$

P.P.S. В каждом zip-архиве есть файл link со ссылкой на первоисточник, откуда был скачан dataset.

Содержание отчета

1. Титульный лист с номером своего варианта, фамилией студента, группы.

2. Теоретическое изложение применяемых алгоритмов.

3. Тексты программ на Python.

4. Числовые данные и графики по результатам работы программ. В объеме достаточном для оценки качества работы исследуемых алгоритмов.

5. Выводы по полученным результатам исследований.