

1. scikit-learn や教科書(Python Machine Learning)における機械学習の流れについて問う

機械学習は(1)データの準備、(2)データの前処理、(3)学習モデルの選択、(4)訓練(学習)、(5)学習モデルによる(未知データに対する)予測、というプロセスがある。

(1) 以下のようにして iris 変数にアヤメのデータセットをセットし、そこから 変数 X と y に値をセットした。

```
iris = datasets.load(iris)
```

```
X = iris.data[:,:] ; y = iris.target
```

これに続いてつぎのようなコードを実行する。

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
```

この X_train, X_test, y_train, y_test 変数はどのような目的のものか、また X と y のサイズがともに 150 行とすると、それぞれのサイズはどうか、なるべく詳しく書け。

(2) データの前処理の一つに「標準化」がある。標準化とはどのようなものか述べよ。また X_train[:,0] を標準化する Python コード、もしくは数式を書け。ここで X_train は Numpy 配列となっており、Numpy の mean と std というメソッドが使えるものとする。

(3) 分類問題を解くことを考える。学習モデルとして、ロジスティック回帰を採用することとした。ロジスティック回帰は sklearn のモジュールにあり、次のようにして使うための準備ができる：

```
from sklearn.linear_model import LogisticRegression
```

ここで lr という変数にロジスティック回帰の学習モデルをセットしたい。ただしパラメタ C の値を 100 とし、考えるパラメタは C だけでよいとする。そのための Python コードを書け。(ヒント: インスタンスを lr に与える)

(4) ここまでの処理により、X_train(標準化済)、y_train, lr には適切な値がセットされているとする。いよいよ学習させたい。そのための Python コードを書け。

(5) 学習させたモデルを X_test (標準化済) に適用したい。そのための Python コードを書け。

2. 誤差平方和(SSE)のコスト関数は次のように定義できる:

$$J(\mathbf{w}) = \sum_i \frac{1}{2} (\phi(z^{(i)}) - y^{(i)})^2$$

ここで $y^{(i)}$ は i 番目のデータに対する正解、そのデータに対し学習モデルの出力が $z^{(i)}$ 、それに活性化関数 $\phi()$ を適用したものが $\phi(z^{(i)})$ である。ここでは ADALINE モデルを学習モデルとして考えおり、 $x_k^{(i)}$ を i 番目のデータの k 番目の要素、 w_k をそれに対する重みとして、 $z^{(i)} = \sum_k (w_k x_k^{(i)})$ と表されるとする。

(1) $\frac{\partial z^{(i)}}{\partial w_k}$ を求めよ。

(2) $\frac{\partial J}{\partial w_k}$ を $y^{(i)}$ 、 $\phi(z^{(i)})$ 、 $x_k^{(i)}$ を用いて表わせ。

3. ロジスティック回帰のコスト関数は $J(\mathbf{w}) = - \sum_i [y^{(i)} \log(\phi(z^{(i)})) + (1 - y^{(i)}) \log(1 - \phi(z^{(i)}))]$ と

表される。ここで $z^{(i)} = \sum_k (w_k x_k^{(i)})$ 、 $\frac{\partial \phi(z)}{\partial z} = \phi(z)(1 - \phi(z))$ である(ただし $\phi()$ としてシグモイド関数を

仮定)ことを用いて、 $\frac{\partial J(\mathbf{w})}{\partial w_k} = - \sum_i (y^{(i)} - \phi(z^{(i)})) x_k^{(i)}$ となることを示せ。