

scikit-learn、Keras、TensorFlow による実践機械学習

東京大学 工学部 3 年 阿部 慧人

2024 年 8 月 18 日

概要

インターンの課題図書の 4 章から 9 章の演習問題を自分で解いたものの答えを載せる。

1 4 章 モデルの訓練

1.1 問題 4.1

数百万個もの特徴量を持つ訓練セットに対して、有効な線形回帰訓練アルゴリズムは、バッチ GD、確率的 GD、ミニバッチ GD である。これらの勾配降下法は特徴量の数 n に対して、 $O(n)$ でしか計算量が増加しない。一方で、正規方程式や SVD は、 $O(n^2)$ 以上の計算量がかかる。

1.2 問題 4.2

特徴量のスケールが大きく異なる場合、勾配降下法を使うと時間がかかるというデメリットがある。これは、損失関数の等高線が平たい形になり直線的に最小値に移動できないことが原因である。

1.3 問題 4.3

ロジスティック回帰モデルの損失関数は、式 (1)(2) である。

$$J(\theta) = -\frac{1}{m} \sum [y^{(i)} \log(p^{(i)}) + (1 - y^{(i)}) \log(1 - p^{(i)})] \quad (1)$$

$$\text{where } p = \sigma(\mathbf{x}^T \theta) \quad (2)$$

この式から、 $J(\theta)$ は、 θ について凸関数とわかるので、極小値は唯一でそれを最小値とすればよい。

1.4 問題 4.4

十分な時間を与えても前提とするモデルの枠組みによっては、同じモデルに収束するとは限らない。具体的には、考えるモデルの損失関数の形による。極小値がただ一つであれば、その極小値に収束して、ただ一つの最適なモデルを得られる。しかし、

極小値が複数あると局所的な解に陥り、必ずしも最適な回にたどり着くとはいえない。

1.5 問題 4.5

バッチ勾配降下法でエポックごとに検証誤差が大きくなっていると考えられるのならば、過学習が生じていると考えることができる。このような時には、検証誤差が最小をとっているところで早期打ち切りをすることが望ましい。

1.6 問題 4.6

ミニバッチ勾配降下法は、毎回の学習ステップで無作為に選んだインスタンスの集合を使って学習する。そのため、その損失関数は確率的な変動を常に含む。よって、検証誤差が少し上昇したとしても、それは確率的な変動によるものである可能性もあるため、過学習だと決めて早期打ち切りをするのは不適切。

1.7 問題 4.7

最も早く最適解近辺に辿り着くのは、確率的勾配降下法かバッチの小さいミニバッチ勾配降下法である。ともに各イテレーションで行う演算が少ないため、短い時間で終わる。一方で、収束はしない。これは、毎回確率的にインスタンスを選ぶことにより損失関数が上下するからである。学習が進むにつれて、学習率を下げれば、収束する。

1.8 問題 4.8

多項式回帰において、訓練誤差と検証誤差の間に大きな差があるというのは、過学習が生じていると考えられることはできる。解決する手法としては、以下の三つが考えられる。1 つが、モデルを正則化

するということ。Ridge 回帰や Lasso 回帰を使う。2 つ目が、多項式モデルの次数を下げること。次数が高いと、自由度が高く学習インスタンスに過剰に追従することがある。3 つ目が、学習インスタンスの数を増やすことである。

1.9 問題 4.9

Ridge 回帰を使っていて、訓練誤差と検証誤差がほとんど同じだが、非常に高い場合、バイアスが高いと考えられる。正則化パラメータ α を下げて、自由度を上げることで、バイアスを下げることができる。

1.10 問題 4.10

線形回帰ではなく、Ridge 回帰を使うべき理由は、線形回帰では自由度が高く過学習を引き起こすことが多いからである。

Ridge 回帰ではなく、Lasso 回帰を使うべき理由は、多くの場合意味のある特徴量が一部であり、Lasso 回帰は意味のある特徴量の係数以外は 0 にしてくれるので、意味のある特徴量のみ抽出できるから。

Lasso 回帰ではなく、Elastic Net を使うべき理由は、訓練インスタンスの数よりも特徴量の数の方が多い時や特徴量間に相関がある時、Lasso では不規則な動きをすることがあるから。

1.11 問題 4.11

2 つのロジスティック回帰分類機を作るべき。屋外・屋内と日中・夜間というクラスは相互排他的ではない。そのため、ソフトマックス回帰分類器のように多クラス出力できないものは役に立たない。

1.12 問題 4.12

Jupyter Notebook に掲載。

2 5 章 サポートベクトルマシン

2.1 問題 5.1

サポートベクトルマシンの基本的な考え方は、二つのクラスの分類する時にその境界として「太さを持った道」を通すことである。この道から最も近いインスタンスとの距離をマージンと呼ぶ。このマージンをできるだけ多くすると、性能が良い分類器になる。その一方で、すべてのインスタンスを完

全に分類する道を考えると、マージンは小さくなる。この妥協点を見つけるのがポイント。

2.2 問題 5.2

サポートベクトルとは、サポートベクトルマシン (SVM) の境界となる「道」の境界と内側に存在するインスタンスのこと。SVM の境界は、このサポートベクトルによってのみ決まり、それ以外のインスタンスは関係ない。

2.3 問題 5.3

SVM を使うときに、入力をスケーリングするのは SVM は特徴量のスケールの影響を受けやすいからである。スケーリングをした方が、決定境界の道が各特徴量を平等に考慮したものが出来上がる。

2.4 問題 5.4

インスタンスの分類するときの確信度を数値化する方法として、決定協会からインスタンスまでの距離をスコアとしたものが考えられる。しかし、これを直接確率に変換する方法はない。scikit-learn では、SVM クラスを使うときに、probability=True とすれば確率をロジスティック回帰によって計算するメソッドが追加される。

2.5 問題 5.5

訓練インスタンスの数 m に対して、計算量は主問題は $O(m)$ で、双対問題は $O(m^2)$ から $O(m^3)$ であるから種問題を解く方がよい。

2.6 問題 5.6

ガウス RBF カーネルを使った SVM で過小適合しているときは、 γ を大きくした方がよい。 γ を大きくすると、ベル曲線の幅が狭くなりより個別のインスタンスに適合するような形状になる。また、ハイパーパラメータ C についても、大きくするとマージン違反に厳しくなるため、よりインスタンスに適合する。

2.7 問題 5.7

ソフトマージン線形 SVM 分類器を QP ソルバーを使って解くことを考える。このとき各変数は以下

のように設定する。

$$H = \begin{pmatrix} 0 & 0 & \dots & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 0 \end{pmatrix} \quad (3)$$

$$\mathbf{f} = (0 \quad \dots \quad 0 \quad c \quad \dots \quad c)^T \quad (4)$$

$$\mathbf{a}'^{(i)} = -t^{(i)} \times [1, \mathbf{x}^{(i)}] \quad (5)$$

を満たす $m \times (n+1)$ 行列 \mathbf{A}' を考えた時、

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}' & I_m \\ \mathbf{O} & -I_m \end{pmatrix} \quad (6)$$

$$\mathbf{b} = (-1 \quad \dots \quad -1 \quad 0 \quad \dots \quad 0)^T \quad (7)$$

このようにすることで、QP の条件式の前半の m 個で、 w に関する条件と後半の m 個で ζ に関する条件が得られる。

2.8 問題 5.8

Jupyter Notebook に掲載。

2.9 問題 5.9

Jupyter Notebook に掲載。

2.10 問題 5.10

Jupyter Notebook に掲載。

3 6章 決定木

3.1 問題 6.1

深さの制限を設けないような決定木を構成するとき、それは Yes と No を繰り返して二分木になることが考えられる。よって、 $\log_2 m \approx 18$ となる。よって深さは、18 ほどだと想定される。

3.2 問題 6.2

一般に親ノードよりも子ノードの方がジニ不純度は小さくなる。これは、CART アルゴリズムでは子ノードのジニ不純度を小さくするように分割されているから。しかし、これは常に成り立つわけではない。というのも、CART アルゴリズムは複数いる子ノードの不純度の加重平均が小さくなるようにしているだけだから。インスタンスを多

く含む子ノードを純粹にすることで、インスタンスの少ない別の子ノードの不純度が親ノードのそれよりも上昇しても、結果としてその加重平均は親ノードよりも小さくなる。

3.3 問題 6.3

過学習をしているときは、正則化ハイパーパラメータのよって条件をより厳しくすることが望まれる。よって、max_depth という上限は下げると良い。

3.4 問題 6.4

入力特徴量の値を増やしても、スケーリングやセンタリングの影響を受けないから、意味はない。

3.5 問題 6.5

計算量は、 $O(n \times m \log_2 m)$ でインスタンスの数 m に対して $m \log_2 m$ に比例する。そのため、インスタンスの数が 10 倍になれば、 $10 \log_2 10m / \log_2 m \approx 12$ 時間になる。

3.6 問題 6.6

インスタンスの数が数千未満のときは、presort することで速度は上がるが、インスタンスの数が多いたときは速度がかなり下がる。

4 7章 アンサンブル学習とランダムフォレスト

4.1 7.1

5 8章 次元削減

5.1 8.1

6 9章 教師なし学習のテクニック

6.1 9.1

参考文献

- [1] オライリー・ジャパン 「scikit-learn、Keras、TensorFlow による実践機械学習」