実践機械学習第４～９章　演習問題(コード問題は別途)

第４章

1.数百万個もの特徴量を持つ訓練セットがあるときに使える線形アルゴリズムは何か

SGD やミニバッチ降下法

2.訓練セットのスケールがまちまちだと悪影響を受けるアルゴリズムは何か。また、対処方法は何か

GD。スケールが大きいものがあるとそれが一番重要視されほかの小さいスケールが意味を持たなくなってしまう。対処法として StandardScaler クラスなどを用いてからアルゴリズムを使用する。

3.ロジスティック回帰を訓練しているときに、GD が局所的な最小値から抜け出せなくなることはあるか

ない。損失関数が凸関数なので、勾配降下法が全体の最小値を見つけられることは保証されている。

4.十分な実行時間を与えれば、すべての GD アルゴリズムは同じモデルに帰着するか

いいえ。損失関数が n 次元で局所的最適解に落ち着く可能性があるため。

5.バッチ勾配降下法をもちいて、エポックごとに検証誤差をプロットしている。検証誤差が絶えず増えているのを気づいた場合、何が起きているのか。対処法は何か。

学習率が高すぎる。学習率を十分に下げることが大事である。

6.検証誤差が上がりだしたときに、ミニバッチ勾配降下法をすぐに中止することはいいことか。

良くない。ミニバッチは最小値を動き回るため検証誤差が毎回下がるとは限らない。

7.最適な解の近辺に最も早く到達するのはどれ。収束するか。他の GD は収束するか。

SGD、ミニバッチ GD。バッチ GD は時間をかければできる。他は 6 のように誤差が上下する。

8.多項式回帰で学習曲線をプロットしたところ、訓練誤差と検証誤差の間に大きな差があった。何が起きているのか。対処方法を 3 つかけ。

過学習を起こしている。自由度を減らす。モデルを正則化。訓練セットを増やす。

9.Ridge 回帰を使っていて、訓練誤差と検証誤差はほぼ同じだが非常に高いことに気づいた。バイアスか分散どちらが高いのか。ハイパラ alpha は上げるべきか下げるべきか。

バイアスが高い。Alpha は下げるべき。

10.・線形回帰でなく、Ridge 回帰を使うべき理由

正則化されたモデルのほうがいい成績を残すため。

・Ridge でなく Lasso 回帰を使うべき理由

自動的に特徴量を選択し、疎なモデルを出力するため。

・Lasso でなく Elastic Net 回帰を使うべき理由

訓練インスタンスの数よりも特徴量の数のほうが多いときや、複数の特徴量の間に大きな相関があるときに不規則な動きをするため、ハイパラは増えるが調整できるほうがよいため。

11.写真を屋外・屋内、日中・夜間で分類したいときソフトマックス回帰分類器か二つのロジスティック回帰分類器を使うべきか。

後者である。他クラス出力できないため。

第 5 章

1.SVM の基本的な考え方

クラスの間にできる限り太い道を通すもの

2.SV とは

「道から外れた」訓練インスタンスではなく、決定境界が決まる、道の際にあるインスタンス

3.SVM を使う際、入力をスケーリングが大事なのはなぜか。

訓練セットがスケーリングされていないとＳＶＭは小さな特徴量を無視してしまうため。

4.SVM 分類器は、インスタンスを分類するとき各進度のスコアが出力できるか。確率はどのくらいか。

直接はできない。しかし、テストインスタンスと決定境界の距離を確信度のスコアとでき、そのクラスに属する推定確率には変換ができない。

5.特徴量が数百個、インスタンスが数百万個の訓練セットで、モデルを訓練するとき、SVM の主問題、双対問題どちらが良いか。

主問題。計算量が断然少ないため

6.RBF カーネル付きの SVM 分類器を訓練する。過少適合してるように見えた場合、γ、C の変化はどうしたらいいか。

正則化がかえって邪魔をしているので、γ かＣを増やして正則化を緩めるべき。

7.出来合いの QP ソルバーを使って、ソフトマージン線形 SVM 分類器の問題を解決する

には QP パラメータをどう設定したらいいか。

ハードマージン問題の QP パラメータを H、f、A、b と呼ぶことにする。ソフトマージン問題の QP パラメータは m 個の追加パラメータと m 個の追加制約を持つ。 ・f は、値がハイパーパラメータ C に等しい m 個の要素を追加した f と等しい。 •b は、値が 0 の m 個の要素を追加した b、と等しい。 •A は、A'の右に m×m の単位行列 Im、その真下に-Im、その他の部分を 0 で埋めたも A' Im のと等しい。

第 6 章

1.百万個のインスタンスを持つ訓練セットで決定木を訓練するとき（無制限で）、おおよその深さはどれぐらいか。

大体 log 2（m）なので、6/0.3010=20 くらいとなる。

2. ノードのジニ不純度は一般に親よりも高いか、それとも低いか。それは一般に高い／低いのか、それとも常に高い／低いのか。

ジニ不純度は一般的に親よりも低い。

3.決定木が訓練セットを過学習している場合、max_depth を下げるとよいか。

よい。

4. 決定木が訓練セットに過小適合している場合、入力特徴量を増やすとよいか。

~~よい。~~

意味がない。決定木は訓練データのスケーリングやセンタリングの影響を受けない。

5.インスタンスが百万個ある訓練セットを対象として決定木を訓練するために 1 時間かかるときに、インスタンスが 1 千万個の訓練セットを対象として別の決定木を訓練するためにどれくらいの時間がかかるか。

10*log(10*1000000)/log(1000000) = 70/6 = 11.67 時間

6.訓練セットのインスタンスが 100,000 個あるとき、presort=True を設定すると訓練のスピードは上がるか。

数千個以上の訓練データを用いているので下がる。

第 7 章

1.　まったく同じ訓練データを使って 5 個の異なるモデルを訓練し、それらがすべての 95%の適合率を達成したとき、それらのモデルを組み合わせたらもっとよい結果が得られる可能性はあるか。もしそうだとすれば、どうすればそのような結果が得られるのか。そうでないとすれば、それはなぜか。

ある。アンサンブルをすれば、平均をとられるため、よりよい結果になる可能性がある。

2. ハード投票分類器とソフト投票分類器の違いは何か。

ハード分類器は、各分類器の予測を集め、多数決で決まったクラス全体の予測であるが、ソフト分類器はすべての分類器がクラスに属する確率を推定できる場合、個々の分類器が推計する確率を平均し、最も確率の高いクラスを予測クラスとして返すという違い。

3. 複数のサーバーで分散処理することによってバギングアンサンプルのスピードを上げることはできるか。ペースティングアンサンプル、ブースティングアンサンブル、ランダムフォレスト、スタッキングアンサンプルではどうか。

バギングアンサンブル：あげられる

ペースティングアンサンブル：あげられる

ブースティングアンサンブル：変わらない

ランダムフォレストアンサンブル：あげられる

~~スタッキングアンサンブル：分からない~~

含まれる予測器が互いに独立しているバギングアンサンブルの訓練は、複数のサーバーで分散処理すればスピードを上げられる可能性がある。同じ理由で、ペースティングアンサンプルやランダムフォレストもスピードが上がる可能性がある。しかし、ブースティングアンサンブルの個々の子測器は前の予測器を基礎として作られるため、逐的な訓練が必要とされ、複数のサーバーの間で訓練を分散処理しても何も得られない。スタッキングアンサンプルの場合、個々の層の予測器はどれも互いに独立しているので、それらを複数のサーバーで並列に訓練することはできる。ただし、ある層の予測器を訓練できるのは、前の層の予測器がすべて訓練されてからである。

4.OOB 検証の長所は何か。

OOB 検証は訓練インスタンスで使用されなかったものを使用するためいちいち別個の検証セットを作らなくてよい

5.Extra-Trees 分類器が通常のランダムフォレストよりも無作為的なのは何によるものか。この余分に無作為的なことにはどのような意味があるか。Extra-Trees は、通常のランダムフォレストと比べて遅いか、それとも速いか。

個々の特徴量の閾値も無作為にしているため。バイアスを少し上げて分散を少し減らす。通常よりも早く訓練できる。

6. 手元のアダブーストアンサンブル訓練データに過小評価している場合、どのハイパーバラメータをどのように調整すべきか。

推定期の数を増やす、ベース推定器を弱く正則化。

学習率も上げるのを検討

7. 勾配ブースティングアンサンブルが訓練セットを過学習している場合、学習率を上げるべきか下げるべきか。

学習率は下げるべきである。

第 8 章

1.データセットを次元削減する主要な理由はなにか。次元削減の主要な欠点は何か。

主要な理由として訓練スピードを上げるだけでなく、データの可視化という点でも役に立つ。大きな欠点は、確実にある程度の情報を失ってしまうという点。

2. 次元の呪いとは何か。

訓練インスタンスごとに数千、数百万もの特徴量を相手にしており訓練が極端に遅くなるだけでなく、良い解を見つけられなくなってしまうということ。

3.データセットを次元削減したあとで、次元を元に戻すことはできるか。できるならどのようにしてするのか。できないならなぜか。

次元を削減している時点で、次元を戻すときの情報がないので完全に元には戻せない。

4. PCA は、高次非線形データセットの次元削減に使えるか。

高次非線形データセットの次元削減に使える。

5.因子寄与率 （explained variance ratio）を 95%に設定して 1,000 次元のデータセットに PCA を 適用する場合、得られるデータセットの次元はどの程度になるか。

因子寄与率の合計が十分な割合になるように次数を選ぶほうが良いため、次元は一概には言えない。

6.通常の PCA、逐次学習型 PCA、ランダム化 PCA、カーネル PCA はどのように使い分けるか。

ランダム化 PCA は完全な特異値分解を使うと計算量は O(n^2*m)+O(n^3)だが、O(m*d^2)+O(d^3)なので d が n よりも小さいと計算が早くなる。逐次学習型 PCA は訓練セットがメモリに収まっていなければできなかったのが、訓練セットミニバッチに分割し一度に 1 つずつミニバッチを渡していくことができる。大規模な訓練セットを相手にするときや PCA をオンライン実行したいときに役立つ。カーネル PCA は暗黙のうちにインスタンスを次の空間にマッピングして、SVM で非線形の分類や回帰を実現するカーネルトリックを応用して、PCA が次元削減のために複雑な非線形射影を実行でき、射影後にもインスタンスのクラスタをうまく保存でき曲がりくねった多様体に近接するデータセットの展開にも使える。

7. データセットに対する次元削減アルゴリズムの性能はどのようにすれば評価できるか。

再構築誤差が小さいと性能は良いといってもよい。これは次元削減をしても大事な情報が残っていることを表しているためである。

8. 2 つの異なる次元削減アルゴリズムを続けて使うことに意味はあるか。

LLE は計算量がおおいため、訓練セットが多いと実行時間が長くなるが LLE を挟む前に、PCA を挟むことによって時間削減が期待できる。

第 9 章

1. あなたならクラスタリングをどのように定義するか。また、クラスタリングアルゴリズムをいくつか挙げなさい。

インスタンスを見分けてクラスタに振り分ける。
クラスタリングアルゴリズム：K 平均法、DBSCAN、凝集クラスタリング、BIRCH、平均法シフト法、アフィニティ伝播法、スペクトラルクラスタリング

2.クラスタリングアルゴリズムの主要な応用分野をいくつか挙げなさい。

顧客セグメンテーション、データ解析、次元削減テクニックとして、異常検知、半教師あり学習の一部として、検索エンジンの一部として、画像セグメンテーション

3.K 平均法を使うときに適切なクラスタ数を選択するための方法を 2 つ説明しなさい。

①クラスタ数 k の関数として慣性をプロットすると、曲線には肘と呼ばれる屈曲点ができることが多く、それが最適解付近であることを示す。
②すべてのインスタンスのシルエット係数の平均であるシルエットスコアと呼ばれるものを使ってよいスコアが最適解付近であることを示している。

4.ラベル伝播とは何か。なぜそのようなものを実装するのか。またどうすれば実装できるか、説明しなさい。

同じクラスタのほかのすべてのインスタンスにも同じラベルをつけること。

~~わからない。~~

すべてのインスタンスに対して K-means 法を用いてクラスタリングし、重心に近いインスタンスのラベルを調べて同じクラスタのラベル無しインスタンスにコピーする。

5.大規模なデータセットに対するスケーラビリティが高いクラスタリングアルゴリズム①を 2 つ挙げなさい。また、高密度の領域を探すクラスタリングアルゴリズム②を 2 つ挙げなさい。
①k-means 法、BIRCH
②DBSCAN、平均シフト法

6.能動学習が役に立つユースケースを挙げなさい。また、どのように実装するか答えなさい。

ラベル労力に見合わないときに用いる。人間の専門家が学習アルゴリズムとやり取りをして、アルゴリズムが特定のインスタンスのラベルを教えてくれと要求してきたときに、ラベルを与えてやる。たとえば、不確実サンプリングがある。

7.異常検知と新規検知の違いは何か。

異常検知は正常なインスタンスとは大きくかけ離れたインスタンスを検知するタスクで、新規検知はアルゴリズムが「クリーン」なデータセットで訓練されていることが前提となっている違い。

8.混合ガウスモデルとは何か。どのようなタスクで使えるか。

パラメータがわからない複数のガウス分布を混ぜたものからインスタンスを生成されていることを前提とする確率的なモデル。一般に楕円体のようなクラスタに分けられる。
異常検知、密度推定、クラスタリング

9.混合ガウスモデルを使っているときに適切なクラスタ数を見つけるための 2 つのテクニックとは何か。

①ベイズ情報量基準(BIC)、赤池情報量基準(AIC)を用いて最小の値が最適解付近である。
②混合ベイズガウスモデルをもちいてアルゴリズムが自動的に最適解を導き出してくれる。