

1.3.2

- ラベルなしのデータ場合
 - 正常時に出現確率が高いなら異常度が低い
- Dの中には異常標本がないか、ごく僅かであると信じられることが必要
- 異常度が高いなら得られる**情報量**は高い
- 異常度が低いなら、得られる情報量は少ない
 - 珍しい観測値を得た方が情報をたくさんゲットできる
- 計測値 x' に対する異常度 $a(x')$ を次のように定義できる
- 「Happy families are all alike; every unhappy family is unhappy in its own way」

1.4 検知器の性能を評価する

異常検知の性能評価の重要な点

- データを分けること
 - 訓練データ(training data)
 - 検証データ(validation data)
 - テストデータ(test data)

交差確認法(cross validation)

- 訓練データで異常検知のモデルを作る
- 検証データでその性能を評価する
- 5分割して4つで訓練し、残りで臆断する
 - これを5回繰り返した平均を性能指標とする

1つ抜き交差確認法(leave-one-out cross validation)

- はずれ値検出の時はNこの標本があるときにN-1個でモデルを作り、残りの1つであたりかはずれかを見る
- N回繰り返す

異常検知の問題点

- (正常標本の数)>>(異常標本の数)
- 楽観的な検出器は異常データを全て見逃す
- 悲観的な検出器は圧倒的多数の正常標本に間違った答えを出す
- 正常標本を使うのか異常標本を使うのかはっきりさせるべき

1.4.1 正常標本精度

正常標本精度(normal sample accuracy)

- 正常標本に対する最も自然な指標
- (正常標本精度) \equiv (実際に正常&正常と判定)/(全ての正常な標本)
- N=100で正常標本が90なら分母は90, 分子には成功した数が入る
- 正答率(detection rate)とも呼ばれる

誤報率(false alarm rate)

- 偽陽性率(false positive rate)とも呼ばれる
- 悪さに注目する
- 陰性を陽性と判断してしまう

1.4.2 異常標本精度

異常標本精度(anomalous sample accuracy)

- (異常標本精度) \equiv (正しく異常と判定した数)/(異常の総数)
- 異常網羅率(cover-age)
- 再現率(recall, リコール)
- ヒット率(hit ratio)
- 真陽性率(true positive ratio)

1.4.3 分岐点精度とF値

- 正答率やヒット率は異常度の閾値の設定で大幅に変わる

異常検出器の性能を表現する

- 図を提示するのが最善の方法
- 分岐点精度(break-even accuracy)
- 性能分岐点での精度のこと
 - 正常標本精度が異常標本精度に一致する点での精度
 - [感想]なんか異常データは少ないからちょっと異常よりにずらした方がたくさん検知できて上手いきそう

F値(F-score)

- 実用的には, 一致する点を求めない

- 正常標本精度 r_0 と異常標本精度 r_1 の**調和平均**(harmonic mean)を閾値やパラメータの異なる値ごとに計算しその最大値を与える点を性能分岐点とするのが便利である

1.4.4 ROC曲線の下部面積

ROC曲線(receiver operating characteristic curve)

- 正常のと異常な標本に分けて精度を閾値の関数として表示する指標の一つ
- **受信者操作特性曲線**とも呼ぶ
- 異常標本精度を誤報率の関数として表した曲線として定義される
- ある閾値 τ に対して、X座標とY座標が
 $(X, Y) = (1 - r_0(\tau), r_1(\tau))$ (1.2)
 のようになる点の集まり
- ROC曲線と横軸が挟む部分を**AUC**と呼び、異常検知器の良さの指標になる

定理1.1(ROC曲線と異常判定閾値の関係)

異常度を式(1.2)で表した時、ROC曲線の傾きの対数は、その点における異常判定の閾値に等しい

証明略

1.5 ネイマン・ピアソン決定則による異常検知の最適性

定理1.2(ネイマン・ピアソンの補題)

- 式(1.2)であたえた異常度の定義は、一定の正常標本精度のもとで異常標本精度を最大にするという意味で最適である。
- 証明の式がわけわからん

2章 ホテリングの T^2 法による異常検知

- 単一の変量正規分布にしたがう独立な標本があたえられたときの古典的な外れ値検出であるホテリングの T^2 法

2.1 多変量正規分布の際尤推定

正規分布(normal distribution)

- ホテリングの T^2 法では、データが異常標本をふくまないか、含んでいても超少ないものとして各標本が独立して次の式に従うと仮定する
- $(x|\mu, \Sigma) \equiv (|\Sigma|^{-1/2}) / (2\pi^{M/2}) * \exp -1/2(x - \mu)^T \Sigma^{-1}(x - \mu)$ (2.1)
- ガウス分布(Gaussian distribution)とも呼ばれる
- **平均と共分散行列**というパラメータがある
- Σ^{-1} を Λ で表して**精度行列**(precision matrix)とよぶ
- **最尤推定**(maximum likelihood)
- 上の二つのパラメータをデータからきめる方法
- N このデータをゲットしたと仮定した対数尤度 L
- $L(\mu, \Sigma|D) = \ln \prod N(x|\mu, \Sigma) = \Sigma \ln N(x|\mu, \Sigma)$ (2.2)
- この式を最大化する μ と Σ を求めるために L を μ で微分して $=0$ とする。
 - この結果は μ となり、相加平均である。

2.2 マハラノビス距離とホテリングの T^2 法

- 最尤推定量を代入することでデータ D を表現する確率密度関数は $p = N(x|\mu, \Sigma)$ のように得られたことになる
- 異常度を定義する
- $a = (x - \mu)^T \Sigma^{-1}(x - \mu)$
- これは観測データがどれだけ標本平均と離れているかの距離を表す
 - **マハラノビス距離**(Mahalanobis distance)とよばれる
- Σ^{-1} は各軸を標準偏差で割るイメージ
 - ばらつきが大きいと、その動きは大目にみる

定理2.1 ホテリングの T^2 法

- M次元正規分布 $N(\mu, \Sigma)$ からのN個の独立標本に基づき μ と Σ を定義。
- 新しく独立標本をみつけたとき、以下が成り立つ。
 1. $x - \mu$ は平均0, 共分散 $N + 1 / N \Sigma$ のM次元正規分布に従う
 2. Σ は $x' - \mu$ と統計的に独立である
 3. $T^2 \equiv (N - M) / ((N + 1)M) * a(x')$ により定義される統計量 T^2 は自由度(M, N-M)のF分布に従う。ただし, $a(x')$ は式(2.9)で定義される
 4. $N \gg M$ の場合は, $a(x')$ は, 近似的に, 自由度M, スケール因子1のカイ2乗分布に従う

ホテリング統計量(Hotelling's statistics)

- 定理2.1の3の T^2 という統計量を呼ぶ
 - 定数倍も含む
- 実用上は, Mがよほど大きくないかぎり $N \gg M$ が成り立つ
 - 最も重要なのは4.である
 - 異常度 a はデータの単位や数値によらず, 常に自由度M, スケール因子1のカイ2乗分布に従うということ
 - スケール因子ってなんだ?

カイ2乗分布(Chi-squared distribution)

- 確率密度関数が

$$\chi^2 = (1/2s\Gamma(k/2))(u/2s)^{(k/2)-1} \exp(-u/2s)$$
 で与えられるもの。
- s: スケール因子
- k: 自由度
- Γ : ガンマ関数

$$\Gamma(z) \equiv \int_0^\infty dt t^{z-1} e^{-t} \quad (2.11)$$
 - 期待値はM, 分散2M
 - 分散がMに比例することは重要
 - 精度良い異常検知のためには, なるべく変数を絞った変数の部分集合ごとに T^2 を計算するのが良い
 - 事前の特徴量の吟味(feature engineering)が必要である所以

アルゴリズム2.1

所与の誤報率 α に基づき, カイ2乗分布から方程式

$$1-\alpha = \int dx \chi^2()$$

により閾値 a_{th} を求めておく

1. 正常標本が圧倒的に多いと考えられるデータから標本平均と共分散行列を求める
2. 新たな観測値 x' について毎回マハラノビス距離を調べる
3. 2.が閾値の a_{th} を超えた場合は警報を出す

ホテリングの T^2 法

- 半導体製造プロセス監視業務をはじめ、広く実世界で使われている
- この手法のパーセント値による閾値はしばしば誤報をもたらす
 - 原因は、F分布やカイ2乗分布における自由度が実際と食い違うことがあるから
 - 解決は、訓練標本に対して計算された異常度に対し改めてカイ2乗分布を当てはめる(7.4節)

2.2節のまとめ

分かったこと

- ホテリング T^2 法は、標本平均から観測値 x' がマハラノビス距離的にどのくらい離れているかで異常かどうかを検知している
- 実際はF分布やカイ2乗分布にそわないこともあり、誤報も少なくない
わからなかったこと
- F分布ってどんな分布だったっけ？
- 共分散行列だから、属性的なやつはいっぱい合っても問題ない？

2.3 正規分布とカイ2乗分布の関係

大事なこと

- ホテリングの T^2 法で最も重要なのは、異常度としてのマハラノビス距離(2.9)が、カイ2乗分布に従うこと

定理2.2 1次元正規変数の平方和の分布

- $N(0, \sigma^2)$ に独立に従う M この確率変数 x_1, \dots, x_M と定数 $c > 0$ により定義される確率変数
 $u \equiv c(x_1^2 + \dots + x_M^2)$

は自由度M,スケール因子 $c\sigma^2$ のカイ 2 乗分布に従う

「カイ 2 乗」

- この定理はなぜカイ 2 乗分布がいかにも x_i^2 を示唆する名前なのかを示している
- マハラノビス距離はM次元の正規変数の 2 次式
 - M個の 2 乗が含まれる
 - マハラノビス距離が自由度Mのカイ 2 乗分布に従うという定理2.1は覚えやすい

2.4 補足:デルタ関数と確率分布の変換公式

- 省略

3章 単純ベイズ法による異常検知

- 多次元のデータに対する異常検知の問題を1次元の異常検知の問題に帰着させるための枠組みを考える
- ラベル付きデータとラベルなしデータに与えた異常度が、多項分布と多次元正規分布に対しどんな式になるか注目して読もう。

3.1 多次元の問題を1次元に帰着する

異常検知の問題を難しくする要素

- 「変数がたくさん合って手に負えない」という状況
- データがM次元の時、 $M=2, M=3$ なら何とかイメージが掴めてもそれ以上は頭で考えるのは難しい

単純ベイズ(naive Bayes)法(ナীবベイズ法)

- 単純ベイズ法はその困難を変数ごとに問題を切り分ける単純な考え方で解決する
- 異常度を計算するために、yが異常かそうでないかによってxの条件付きの分布を与える
- それに含まれるパラメータをデータから決める
- 単純ベイズ法のモデルを仮定

$$p(X|y) = p(x_1|y) \dots p(x_M|y) = \prod_{i=1}^M p(x_i|y)$$

- M次元のそれぞれが統計的に独立であることを示す

統計的に独立である

- 最尤推定のための対数尤度の式を書き下してみるとわかりやすい
- x の条件付き分布が $p(x_i|\theta_i^y, y)$ のように未知パラメータを含む形で書けるとする
- 迷惑メール分類に使う多項分布であれば、 θ は i 番目の語の出現確率そのもの
- y は0番(普通メール)と1番(迷惑メール)の出現頻度は異なるので、その違いを区別する
- この時の対数尤度

$$L(\Theta|\mathcal{D}) = \sum \sum \ln p(x_i^{(n)}|\theta_i^{y^{(n)}}, y^{(n)})$$

結局はシータと y の時の x である条件付き確率の対数を示している

- \mathcal{D}^0 は $y = 0$ = 正常となる標本の集合
- \mathcal{D}^1 は $y = 1$ = 異常となる標本の集合
- n に関する和は、それぞれの集合の要素に対してとる
- Θ : 異なる i と y に対するパラメータ θ_i^y を全部まとめて表記した記号
- θ_i^1 の最尤解を与える条件式
$$0 = \partial L / \partial \theta_i^1 = \partial / \partial \theta_i^1 * \sum \ln p(x_i^{(n)}|\theta_i^1, y = 1)$$
- 最右辺には添字 i と $y=0$ に対応する項しか寄与しない
- 問題が変数ごと、 y ごとに切り分けられたということ

定理3.1 変数が統計的に独立な場合の最尤推定

式(3.1)のように変数ごとに積の形となっている場合、 M 変数のそれぞれに対して別々に最尤推定することで、モデルのパラメータを求めることができる。

$$p(X|y) = p(x_1|y) \dots p(x_M|y) = \prod_{i=1}^M p(x_i|y) \quad (3.1)$$

名前の由来

- 「ベイズ」は変数同士を独立とみなすという上記の「単純な」想定に、ベイズ決定則と呼ばれる分類規則を併用して分類器が構築されるのが通例だから

3.1 まとめ

- 単純ベイズ法は、「各変数が独立だとみなすモデリング手法を異常度(式1.2)に適用したもの」である。

3.2 独立変数モデルの元でのホテリング T^2 法

変数の独立性を仮定するモデル

- ラベルなしデータでも適用できる
- ホテリングの T^2 法
 - $T^2 = (N - M)/(N + 1)M * a(x')$ で表される統計量
 - F分布に従う
 - F分布は、分散が等しいかを調べるときなどに使われる
- データは共分散行列の対角成分のみ取り出したもの
 - $p(x) = \Pi \mathcal{N}(x|\mu, \Sigma)$
- 定義3.1は成り立つので、対数尤度を微分して最尤解が求められる
 - 変数ごとに積の形になっており、別々に最尤推定することでモデルのパラメータを求められる
- $\mu = 1/N \sum x, \Sigma = 1/N \sum (x - \mu)^2$ (3.4)