## 1.3.2

- ラベルなしのデータ場合
  - 正常時に出現確率が大き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，残りで腱鞘する
  - 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)≡(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)≡(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くいきそう

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(τ),r_1(τ))　(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

- 単一の多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
- $(x|\mu,\Sigma) \equiv (|\Sigma|^{-1/2})/(2\pi^{M/2}) * exp{-1/2}(x-\mu)^T\Sigma^{-1}(x-\mu) \quad (2.1)$
- 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(μ,Σ|D)=lnΠN(x|μ,Σ)=Σln N(x|μ,Σ)　(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|μ,Σ)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
- $\Sigma^{-1}$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(μ,Σ)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-μは平均0, 共分散$N+1/N\Sigma$のM次元正規分布に従う
  2. $\sum$はx'-μ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 statictics)

- 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>>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
  $\Gamma(z) \equiv \oint[0,\infty]dt\ t^{z-1}e^{-t}$ (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-α=∮dx χ^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, σ^2)に独立に従うMこの確率変数x1,...,x_Mと定数c>0により定義される確率変数
  u≡c(x_1$^{2+...+x}$_M2)

は自由度M,スケール因子cσ^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(**ナイーブベイズ**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
  $p(X|y) = p(x_1|y)...p(x_M|y) = \mathit{\Pi}[i = 1, M]p(x_i|y)$
- 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$\theta_i^y$を全部まとめて表記した記号
- $\theta_i^1$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。

$$p(X|y) = p(x_1|y)...p(x_M|y) = \Pi[i = 1, M]p(x_i|y) \quad (3.1)$$

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
  - $p(x) = \Pi\mathcal{N}(x|\mu, \sum)$
- 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
- $\mu = 1/N \sum x, \sum = 1/N \sum (x-\mu)^2$ (3.4)

- これを使って異常度(2.9)を求める
  $a(x') = \sum((x'_i - \hat{\mu_i})/\hat{\sigma}_i)^2$ (3.5)
- M次元ベクトルとしての観測値x'の異常度は，M個の変数のそれぞれに対して計算された異常度の和
  - 変数同士が独立であれば有用である

変数間の相関と異常度のつながり

- 式(3.5)は変数間の相関がどのように異常判定に関係するかを理解する上で有用
- 典型的な二つの状況(図3.1)
  - 図の赤い四角に入っていれば正常で外は異常と判定する
  - 線形相関がある場合は，異常判定の枠が不当に大きくなる
    - 変数個々に見ると左上や右下の異常判定を見落としてしまう
  - M変数を2つずつ組にして見るという方法がある

# 3.2 まとめ

- 変数の独立性を仮定するモデルは，ラベルなしデータにも適用できる
- ホテリングの$T^2$法でも変数が独立したものとして考えることができる
- 各変数で異常度を出し，それらの総和で異常を出すことができる

# 3.3 多項分布による単純ベイズ分類

- 単純ベイズ法を多項分布について適用
- 迷惑メールフィルターの基本手法

## 3.3.1 多項分布: 頻度についての分布

- 「頻度」についてのデータは重要
  - 通販サイトの管理者はどの商品が何回閲覧されたかに興味がある
  - 図書館の司書はどのジャンルの本が何冊貸し出されたかに興味がある
  - 頻度専用の分布を使った方が正確かつ妥当な分析ができる

迷惑メールの振り分け

- 迷惑メールフィルターは，それぞれのメールにおける単語の頻度を計算し特徴量として使用
- 使われそうな単語をあらかじめ辞書登録しておき，ベクトル化する
  $x = (0, 0, 0, 1, 0, 0, 2, 0, ...)^T$のように表される
- **単語の袋詰め**(bag-of-words)
  - 考えられる単語を集めてベクトルとして展開したもの
    $Mult(x|\theta) = (x_1 + ... + x_M)!/(x_1!x_2!...x_M!) * \theta_1^{x_1}...\theta_M^{x_M}$ (3.6)
- このようなxの分布を**多項分布**と呼ぶ
  - $M = 2$の時は2項分布
    - 頻度を表現するための自然な分布

多項分布の特徴

- 1つのメールに現れる単語の総数が決まれば分布が積の形になり，各$x_i$ごとに分離している

## 3.3.2 多項分布の最尤推定

迷惑メールの分類問題

- 過去のメールN通が正常y=0,異常ｙ=1のデータとともに蓄積している場合を考える
- 異常度を求めることで判定をする

最初のステップ

- yごとにxの分布を求める
- y=0(正常)とy=1(異常)に対応して2つのモデルを仮定する
  - Mult$(x|\theta^0)$
  - Mult$(x|\theta^1)$
- 未知パラメータ$\theta^0$と$\theta^1$を最尤推定する

- モデルパラメータの対数尤度
  $L(\theta^0, \theta^1|\mathcal{D}) = \sum\sum x_i^{(n)} ln\, \theta_i^1 + \sum\sum x_i^{(n)} ln\, \theta_i^0 +$ (定数) (3.7)
- (定数)は未知パラメータ$\theta_0, \theta_1$に関係しない定数
- Lを制約(θの和が1)のもとで最大化することが問題
  - $\sum \theta_i^1 = 1$および$\sum \theta_i^0 = 1$
- 制約をラグランジュ乗数で取り込むと，θ_0の第i成分についての最適解の条件は次のようになる
  - $0 = \partial / \partial \theta_i^0 (L - \lambda^0 \sum \theta_j^0 - \lambda^1 \sum \theta_j^1) = 1/\theta_i^0\ (\sum x_i^{(n)} - \lambda^0)$
- これより，θについて解が得られる
  - $\hat{\theta_i^0} = (\sum x_i^{(n)}) / (\sum\sum x_j^{(n)}) =$
    ($D^0$における単語$i$の出現総数)/($D^0$における全単語の出現総数)
    - y=1の時も同様

**スムージング**(smoothing)

- 一度も出てこない単語のパラメータが0になることを防ぐ

## 3.3.3 迷惑メールの分類

分類する

- 未知のメールの単語袋詰め表現$x'$がきたとして，迷惑メール（y'=1）か普通メール(y'=0)かを分類する

判定スコアの式

- 異常度の式(1.2)でp()が多項分布に対応しているので，代入して整理する
  $a(x') = \sum x_i' ln\ (\theta^1 / \theta^0)$ (3.9)