テキストデータの統計科学入門

9章　テキストと情報量

茨城大学大学院理工学研究科情報工学専攻

12NM712L　小幡智裕

尤度

- 尤度
  - 標本データが、ある母集団から得られる確率のこと
  - 人為的に作成した複数のモデルの候補の中で、どのモデルが真のモデルに近いかを判断する必要がある。その際に尤度が用いられ、尤度が高いモデルを標本データに最も適していると判断する
  - 尤度を求めるには尤度関数が必要である

尤度関数

確率変数 $X = \{x_1, x_2, ..., x_n\}$、パラメータ $\theta = \{\theta_1, \theta_2, ..., \theta_p\}$ の関数 $f(X, \theta)$ があるとする.この関数はθを変数とする関数であり、その同時確率密度関数は

$$L(\theta) = f(\theta, x_1) f(\theta, x_2) \dots f(\theta, x_n) = \prod_{i=1}^{n} f(\theta, x_i)$$

で表わされ、この関数を尤度関数とよぶ

最尤推定

尤度$L(\theta)$を最大とする$\hat{\theta}$をθの推測値とするとその表記は

$$\hat{\theta} = \arg\max L(\theta)$$

となり、これを最尤法あるいは最尤推定法とよぶ

実際の計算では対数尤度関数を用いる

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log(\, f(\theta, x_i))$$

二項分布の最尤推定量の導出例

二項分布の尤度関数は

$$L(\theta)={}_nC_x\theta^x(1-\theta)^{n-x}$$

であり対数を尤度関数は

$$l(\theta)=\log({}_nC_x)+x\log(\theta)+(n-x)$$

θを変数とした上式の偏微分を0にすると

$$\frac{\partial l(\theta)}{\partial\theta}=0+\frac{x}{\theta}-\frac{n-x}{1-\theta}=0$$

となり、解は $\theta=\frac{x}{n}$ である

情報量

- シャノンの情報量（エントロピー）
  - 記号列おける平均統計量
  - 1記号あたりのばらつきに関する情報量である
  - エントロピーが小さいほどデータのばらつきが大きい

定義

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

$0 \leq H(X) \leq \log(n)$ という性質がある

結合エントロピー

2つの確率変数をX,Yで表し、x_i, y_k が同時に現れる確率が

$p(x_i, y_k)$ であるとき

$$H(X,Y) = -\sum_i \sum_k p(x_i, y_k) \log p(x_i, y_k)$$

を結合エントロピーという

$H(X,Y) \leq H(X) + H(Y)$ という性質をもっている

条件付きエントロピー

条件付き確率に対応するエントロピーを条件付きエントロピーとよび、次の式で定義される

$$H(Y \mid X) = -\sum_i \sum_k p(x_i)\, p(y_k \mid x_i) \log p(y_k \mid x_i)$$
$$= -\sum_i \sum_k p(x_i, y_i) \log p(y_k \mid x_i)$$

条件付きエントロピーは

$$H(X,Y) = H(Y) + H(X \mid Y)$$
$$= H(X) + H(Y \mid X)$$

という性質を持っている

相対エントロピー

確率変数Xの2つの確率分布p,qが与えられたとき

$$KLD(p \| q) = \sum_i p(x_i) \log \frac{p(x_i)}{q(x_i)}$$

相対エントロピーあるいはカルバックライブラーダイバージェンスとよぶ

KLDは次の性質を持っている

$$KLD(p \| q) \geq 0$$

$$KLD(p \| q) = 0 \Leftrightarrow p = q$$

$$KLD(p \| q) \neq KLD(q \| p)$$

相対エントロピー

KLDは非対称であるため、拡張された距離測度が提案されている

$$d_{IR}(p,q)=\sum_i\left[p(x_i)\log\frac{2p(x_i)}{p(x_i)+q(x_i)}+q(x_i)\log\frac{2q(x_i)}{p(x_i)+q(x_i)}\right]$$

次の性質を持っている

$$d_{IR}(p,q)\geq 0$$

$$d_{IR}(p,q)=0\Leftrightarrow p=q$$

$$d_{IR}(p,q)=d_{IR}(q,p)$$

相互情報量

2つの確率変数X,Yの間の相互情報量は、次の式で定義される

$$I(X;Y) = KLD\left[p(x,y) \parallel p(x)p(y)\right]$$

$$= \sum_i \sum_k p(x_i, y_k) \log \frac{p(x_i, y_k)}{p(x_i)p(y_k)}$$

相互情報量とエントロピーには以下の関係がある

$$I(X,Y) = H(X) - H(X \mid Y)$$

$$= H(Y) - H(Y \mid X)$$

$$= H(X) + H(Y) - H(X,Y)$$

相互情報量の値が大きいほど、相対的に多く組み合わせて使用されていることを意味する

情報量規準

作成したモデルを評価するための規準となる量を一般的に情報量規準と呼ぶ

- AIC、BIC、EIC、TIC、GICなどがある

AIC=-2(最大対数尤度)+2(モデルに含まれた独立なパラメータの数)

モデル評価の例

- 三島由紀夫の4作品の文節の長さの分布のモデルの例を示す
- 文を文節ごとに切り分け、各文節が何文字により構成されているかを調べたデータの実測値を何らかの確率分布でモデリングすることを考える
- ここではポアソン分布(M1),対数正規分布(M2)を用いてモデリングすることを考える
- 実測値と2つのモデルによる推測値のデータを表1に示す

表1

文字数	実測値		ポアソンモデル		対数正規モデル	
	F	p	F(M1)	p(M1)	F(M2)	p(M2)
1	53	0.006	776	0.084	47	0.005
2	1988	0.216	1919	0.209	2043	0.222
3	3538	0.385	2372	0.258	3317	0.361
4	1927	0.209	1956	0.213	2163	0.235
5	999	0.109	1209	0.131	999	0.109
6	409	0.044	598	0.065	399	0.043
7	189	0.021	247	0.027	150	0.016
8	53	0.006	87	0.009	56	0.006
9	34	0.004	27	0.003	21	0.002
10~	10	0.001	7	0.001	8	0.001
合計	9200	1	9198	1	9203	1

モデル評価(AIC)

- 推測値から両モデルの対数尤度を求める

$$l(M1) = 53 \times \log(0.084) + 1988 \times \log(0.209) + \\ \ldots + 34 \times \log(0.003) + 10 \times \log(0.001) = -15372.23$$

$$l(M2) = 53 \times \log(0.005) + 1988 \times \log(0.222) + \\ \ldots + 34 \times \log(0.002) + 10 \times \log(0.001) = -14495.93$$

- 対数尤度を用いてAICを求める

$$AIC(M1) = -2 \times (-15372.23) + 2 \times 1 = 30746.46$$

$$AIC(M2) = -2 \times (-14495.93) + 2 \times 2 = 28995.85$$

- AIC(M2)<AIC(M1)より、対数正規分布のほうがポアソン分布より真の分布に近いと判断される

Rで作成した折れ線グラフ