言語処理のための機械学習入門
1.5〜1.6

10NM733X 林華

1.5 パラメータ推定法

- **データが与えられたとする**
  - このデータある種類の確率分布
  - パラメータの値が知らない

- **データからパラメータの値を推定**
  - 最尤推定
  - 最大事後推定

1.5.1 i.i.d. と 尤度

i.i.d. ：独立に同一の確率分布に従うデータのこと

- **サンプルデータ** D = $\{ x^{(1)},..., x^{(N)} \}$
- **尤度**:生成確率P(D) $P(D) = \prod_{x^{(i)} \in D} p(x^{(i)})$
- **対数尤度**:logP(D)

$$\log P(D) = \log \prod_{x^{(i)} \in D} p(x^{(i)}) = \sum_{x^{(i)} \in D} \log p(x^{(i)})$$

$$\log P(D) = \sum_{x} n_x \log p(x)$$

1.5.2 最尤推定

- **最尤推定**：対数尤度が最も高くなるようにパラメータを決定する方法であり、「できる限りデータにフィットさせる」推定方法
- 最適値を求めるには**微分**と**ラグランジュ法**の利用
  - ラグランジュ法を制約条件付きの場合に使う

1.5.3 最大事後確率推定ーその1

- **事前確率分布**：パラメータθの確率分布が分かっている。つまり、パラメータがどんな値を取るかが事前にわかる場合
  - 例：ポアソン分布の平均μが10に近い値を取りやすい
- **事後確率分布**：データDが与えられたときのパラメータθの確率分布P(θ|D)
- **最大事後確率推定**：事後確率P(θ|D)が最大になるようにパラメータを決定

$$(1) \arg\max_{\theta} P(\theta \mid D) = \arg\max_{\theta} \frac{P(\theta) \cdot P(\theta \mid D)}{P(D)}$$

$$= \arg\max_{\theta} P(\theta) \cdot P(\theta \mid D)$$

$$(2) \log P(\theta) \cdot P(\theta \mid D) = \log P(\theta) + \sum_{x^{(i)} \in D} \log P(x^{(i)} \mid \theta)$$

1.5.3 最大事後確率推定ーその2

比較

最尤推定: $\sum_{w} n_w \log p_w$

MAP推定: $\log P(\mathrm{p}) + \sum_{w} n_w \log p_w$

MAPが制約条件付きなので**ラグランジュ乗数** を利用

$$(1) L(\mathrm{p}, \lambda) = \log P(\mathrm{p}) + \sum_{w} n_w \log p_w + \lambda \left(\sum_{w} p_w - 1 \right)$$

$$(2) \frac{\partial L(\mathrm{p}, \lambda)}{\partial p_w} = \frac{1}{p_w} + \frac{n_w}{p_w} + \lambda$$

$$(3) \sum_{w} p_w = 1$$

$$(4) p_w = \frac{n_w + 1}{\sum_{w} (n_w + 1)} = \frac{n_w + 1}{\sum_{w} (n_w + |V'|)}$$

1.6 情報理論

- **エントロピー**
- **カルバック・ライブラー・ダイバージェンス**
- **ジェンセン・ジャノン・ダイバージェンス**
- **自己相互情報量**

1.6.1 エントロピー

- **エントロピー**：確率変数の乱雑さを測るための尺度
  - 高いほど、集中性がある
  - 注意点：「データ分布のエントロピー」だが、「データのエントロピー」ではない
- **定義**：$H(P) = \sum_x -P(X=x)\log P(X=x)$
- **条件付きエントロピー**

$$(1) H(X|Y=y) = -\sum_x P(X=x|Y=y)\log P(X=x|Y=y)$$

$$(2) H(X|Y) = \sum_y P(Y=y)H(X|y)$$

$$= -\sum_y P(Y=y)\sum_x P(X=x|Y=y)\log P(X=x|Y=y)$$

$$= -\sum_{x,y} P(X=x,Y=y)\log P(X=x|Y=y)$$

1.6.2 Kullback-Leibler divergence

- **KLダイバージェンス**：二つの確率分布に対して、両者間の異なり具合を測るものである。
- **定義** $D_{KL}(P\|Q) = \sum_x P(X=x)\log\frac{P(X=x)}{Q(X=x)}$
- **注意点**：
  - 「データ間のKLダイバージェンス」という表現は意味をなさない
  - 同じ事象空間上の確率分布であること
  - 厳密な距離ではない

1.6.3 Jensen-Shannon divergence

- **JSダイバージェンス**：KLダイバージェンスの平均である
- **定義**

$$D_{JS}(P\|Q) = \frac{1}{2}(D_{KL}(P\|R) + D_{KL}(Q\|R))$$

$$= \frac{1}{2}\left(\sum_x P(x)\log\frac{P(x)}{R(x)} + \sum_x Q(x)\log\frac{Q(x)}{R(x)}\right)$$

$$= \frac{1}{2}\left(\sum_x P(x)\log\frac{P(x)}{\frac{P(x)+Q(x)}{2}} + \sum_x Q(x)\log\frac{Q(x)}{\frac{P(x)+Q(x)}{2}}\right)$$

- **改善点**：KLダイバージェンスの問題点が存在しない。0log0=0

1.6.4 自己相互情報量

- **自己相互情報量**：二つの確率変数が関連している度合いを測るものである。
- **定義** $PMI(x,y) = \log\frac{P(x,y)}{P(x)P(y)}$
- **解釈**：xとyが共起関係が高い場合、P(x,y) > P(x)P(y)、よってPMI(x,y) > 0。一方、共起関係薄い場合、PMI(x,y)<0。