ベイズ推論による機械学習入門

ベイズ推論による学習と予測(3.1-3.2)

2022/5/11 情報理工学科 竹川修平

目次

第3章 ベイズ推論による学習と予測 (3.1-3.2)

3.1 学習と予測

学習と予測

- 一般的に機械学習の分野では, モデルの持つパラメータの値をデータから決定することを**学習 (training, learning)**という.
- ベイズ推論の枠組みでは, パラメータも不確実性を伴う確率変数として扱うので, 確率計算によってデータを観測した後のパラメータの事後分布を求めることが**学習**にあたる.
- 多くの場合では単純にパラメータを得るだけでなく, まだ観測されていない値に関する予測を行うことも主要な課題になる. 予測分布に関しても確率推論を使って求め, 未知の値に対する平均値やばらつき具合などの各種期待値を調べたり, サンプルを得ることによって視覚的に予測を理解することが行われる.

パラメータの事後分布

-

$\mathcal{D}$: 訓練データの集合

-

θ: モデルに含まれる未知のパラメータ

ベイズ学習では次のような同時分布$p(\mathcal{D},\theta)$を考えることでデータを表現するモデルを構築
$$p(\mathcal{D},\theta) = p(\mathcal{D}\vert\theta)p(\theta)$$

-

パラメータに関する不確実性は事前分布$p(\theta)$を設定することで反映される.

-

$p(\mathcal{D}\vert\theta)$は特定のパラメータ$\theta$からどのように$\mathcal{D}$が発生したかを記述しており, これを$\theta$の関数とした場合は**尤度関数(likelihood function)**と呼ばれる.

データ$\mathcal{D}$を観測した後ではパラメータの不確実性は次のように更新される $$ p(\theta\vert\mathcal D) = \frac{p(\mathcal D\vert\theta)p(\theta)}{p(\mathcal D)}$$

-

この条件付き分布$p(\theta\vert\mathcal D)$を計算することがベイズ学習の枠組みでの「学習」にあたる

-

事後分布$p(\theta\vert\mathcal D)$は$p(\theta)$と比べて尤度関数 $p(\mathcal{D}\vert\theta)$ を通すことによって観測データ$\mathcal D$に関する特徴を捉えていることが期待できる

3.1.2 予測分布

-

x_*: 未観測のデータ

学習されたパラメータの分布を使って**予測分布(predivtive distribution)**を計算することによって未観測データ$x_$*に関する知見を得られる. $$p(x*\vert\mathcal D) = \int p(x_*\vert\theta)p(\theta\vert\mathcal D)d\theta $$

-

$p(\theta\vert\mathcal D)$で様々なθについて重み付けして$p(x_*\vert\mathcal D)$を平均化

-

上のモデルではデータ$\mathcal D$も未観測値$x_$*もパラメータ$*|theta$*に発生過程が支配されているが, $*|mathcal D$*と$x_$*には直接的な依存関係は仮定おらず, パラメータが与えられたもとで条件付き独立であるといえる. このような仮定を置くとき, 観測データは i.i.d.(independent and identically distributed)であるという.

このような仮定において, 同時分布は次のようになる $$p(\mathcal D,x_*|*theta) = p(|*mathcal D|*vert|*theta)p(x*\vert\theta)p(\theta)$$ データ$\mathcal D$が手元にあるとすれば, 残りの変数の事後分布は

$$ \begin{aligned} p(x_, |*theta|*vert|*mathcal D)&=|*frac{p(|*mathcal D, x_|*ast, |*theta)}{p(|*mathcal D)} | &=|*frac{p(|*mathcal D|*vert|*theta)p(x*\vert\theta)p(\theta)}{p(\mathcal D)} \ &=p(x*\vert\theta)p(\theta\vert\mathcal D) \end{aligned} $$

-

上の式を, 周辺化によってθを除去すれば予測分布が得られる.

データ$\mathcal D$を観測していない状態で予測分布を求めても良い. $$p(x_*) = \int p(x_*\vert\theta)p(\theta)d\theta $$

-

この場合は事前知識$p(\theta)$だけに頼っているので非常に大雑把な予測である.

データ$\mathcal D$と未観測値$x_\ast$が条件付き独立なモデルにおいて, 未観測値$x_\ast$を予測するには, まず, $\mathcal D$を用いてパラメータの事後分布を学習し, その時点で$\mathcal D$は捨て, 得られた事後分布を周辺化することによって予測分布を求める.

-

データ$\mathcal D$の情報を全て事後分布$p(\mathcal D\vert\theta)$に押し込めるのは計算上非常に便利だが, データ量が変化しても, モデルの表現能力は変化しないという制限がある.

-

データ$\mathcal D$に合わせて予測モデルの表現能力を柔軟に変える確率モデルとして**ガウス過程(Gaussian process)**や**ベイジアンノンパラメトリクス(Bayesian nonparametrics)**がある.

3.1.3 共役事前分布

共役事前分布

共役事前分布(conjugate prior)

事前分布$p(\theta)$と事後分布$p(\theta\vert\mathcal D)$が同じ種類の確率分布を持つように設定された事前分布.

-

どのような事前分布が共役になりうるかは尤度関数$p(\theta\vert\mathcal D)$の設計による

- ガウス分布のようにパラメータを二つもつような分布では, どのパラメータを学習させたいかによって共役事前分布が異なる.

共役事前分布を使う利点

事後分布や予測分布の計算が簡単かつ効率的にできる.

データセット$\mathcal D_1$を観測したあとの事後分布は次のようになる. $$p(\theta\vert\mathcal D_1)\propto p(\mathcal D_1\vert\theta)p(\theta)$$ さらに新規データセット$\mathcal D_2$を観測した場合の事後分布は次のようになる. $$p(\theta\vert\mathcal D_1, \mathcal D_2)\propto p(\mathcal D_2\vert\theta)p(\theta\vert\mathcal D_1)$$

-

上式のようにデータセットを小分けにして逐次的に学習する場合, 共役事前分布を用いれば$p(\theta)$, $p(\mathcal D_1\vert\theta)$, $p(\mathcal D_2\vert\theta)$が全て同じ形式になりプログラムによる実装がシンプルになる.

- 解析的に事後分布を求めることができない複雑な拡張モデルにおいて, 共役な分布を部分的に組み合わせるて全体のモデルの構築をすることで計算効率の高い近似アルゴリズを導ける(4章以降).

尤度関数	パラメータ	共役事前分布	予測分布
ベルヌーイ分布	μ	ベータ分布	ベルヌーイ分布
二項分布	μ	ベータ分布	ベータ・二項分布
カテゴリ分布	$\bm{\pi}$	ディリクレ分布	カテゴリ分布
多項分布	$\bm{\pi}$	ディリクレ分布	ディリクレ・多項分布
ポアソン分布	λ	ガンマ分布	負の二項分布
1次元ガウス分布	μ	1次元ガウス分布	1次元ガウス分布
1次元ガウス分布	λ	ガンマ分布	1次元ステューデントのt分布
1次元ガウス分布	μ, λ	ガウス・ガンマ分布	1次元ステューデントのt分布
多次元ガウス分布	$\bm{\mu}$	多次元ガウス分布	多次元ガウス分布
多次元ガウス分布	$\bm{\Lambda}$	ウィシャート分布	1次元ステューデントのt分布
多次元ガウス分布	$\bm{\mu},\bm{\Lambda}$	ガウス・ウィシャート分布	1次元ステューデントのt分布

- **負の二項分布(negative binominal distribution)**や**スチューデントのt分布(student's t distribution)**は以降の予測分布の計算で登場する.

共役でない事後分布の利用

- 尤度関数に対応する共役事前分布をそのまま使うとデータに関する興味深い構造をうまく捉えられないケースがある. そのような場合は共役でない事前分布を使うこともある.
- 例としてガウス分布の平均パラメータに対して共役でないガンマ分布を過程すると事前分布はガンマ分布にならない. このような場合**MCMC(Markov chain Monte Carlo)**や**変分推論(variational inference)**を使うアイデアがある.

変分推論

η:変分パラメータ(variational parameter)を使った分布$q(\theta;\eta)$で事後分布$p(\theta\vert\mathcal D)$を近似的に表現できると仮定し, 以下の最小化問題を解く. $$\eta_{opt} =\underset{\eta}{\text{argmin}}\text{KL}[q(\theta;\eta)\vert\vert p(\theta|\mathcal D)]$$

- 通常この最小化問題は解析的に解くことができないので, **勾配法(gradient method)**などの最適化アルゴリズムを使う.
- 共役事前分布を使った解析的な計算と比べて, 最適化のための計算コストが余分にかかる.

-

得られた近似分布$q(\theta;\eta_{opt})$がどれだけ事後分布を近似できているかは一般的には把握できない. ただし, 複数の近似分布の仮定{$q(\theta;\eta_1),...,q(\theta;\eta_K)$}でどれが最もよいかは**ELBO(evidence lower bound)**という値を使って定量的に評価できる.

3.2.1 ベルヌーイ分布の学習と予測

ベルヌーイ分布の学習と予測

$x\in{0,1}$上の確率分布であるベルヌーイ分布のパラメータμの分布の推論を考える.
$$p(x\vert\mu)=\text{Bern}(x\vert\mu)\coloneqq\mu^x(1-\mu)^{1-x}$$ ベルヌーイ分布のパラメータの要件$\mu\in(0,1)$を満たす値を生成してくれるベータ分布を事前分布として採用する. $$p(\mu)=\text{Beta}(\mu\vert a,b)\coloneqq C_B(a,b)\mu^{1-a}(1-\mu)^{1-b}$$

-

ここでa,bは事前分布$p(\mu)$をコントロールするためにパラメータになっている. μ自体がパラメータなのでa,bをパタメータのためのパラメータということで**超パラメータ(hyper-parameter)**と呼ばれる.

- 今回のモデルでは超パラメータの学習は行わず, 既知の値として与えられているとする.

事後分布を実際に計算してみる.

-

$\bm{X}={x_1, ..., x_N}$: N個のデータ点

$$ \begin{aligned} p(\mu\vert\bm X)&=\frac{p(\bm X\vert\mu)p(\mu)}{p(\bm X)} \ &= \frac{{\prod_{n=1}^{N}p(x_n\vert\mu)}p(\mu)}{p(\bm X)}\ &\propto{\prod_{n=1}^{N}p(x_n\vert\mu)}p(\mu)\ \end{aligned} $$ ここで対数をとる $$ \begin{aligned} \text{ln}p(\mu\vert\bm X)&=\sum_{n=1}^{N}\text{ln}(x_n\vert\mu)+\text{ln}p(\mu)+\text{const.}\ \end{aligned} $$

$$\begin{aligned} &=\sum_{n=1}^{N}x_n\text{ln}\mu+\sum_{n=1}^{N}(1-x_n)\text{ln}(1-\mu)\ &\quad+(a-1)\text{ln}\mu+(b-1)\text{ln}(1-\mu)+\text{const.}\ &=(\sum_{n=1}^{N}x_n+a-1)\text{ln}\mu+(N-\sum_{n=1}^{N}x_n+b-1)\text{ln}(1-\mu)+\text{const.} \end{aligned}$$

<参考>ベータ分布(対数)

$$\text{lnBeta}(\mu\vert a,b)=(a-1)\text{ln}(\mu)+(b-1)\text{ln}(1-\mu)+\text{ln}C_B(a,b)$$

$$\hat a=\sum_{n=1}^{N}x_n+a, \hat b=N-\sum_{n=1}^{N}x_n+b$$ とおけば

$$ p(\mu\vert\bm X) = \text{Beta}(\mu\vert\hat a,\hat b)$$

- 事後分布が事前分布と同じ形式(ベータ分布)となった.

- $\hat a$はaに$x=1$となる回数が追加され, $\hat b$はbに$x=0$となる回数が追加されている. コインで例えると, ベータ分布は今までに表と裏が何回ずつ出たかを記憶する役割を果たしている.

(参考) 経験ベイズ法

3.2.1節では事後分布を求めた. この際パラメータ$\hat a, \hat b$を求めたが, これは事前分布のパラメータa,b自体を更新したわけではない.a,bなどの超パラメータ自体を観測データに合わせて直接調整する方法も存在し, これは**経験ベイズ法(empirical Bayes)**と呼ばれている.確率推論によって導かれる手法ではないため, 厳密にはベイズ手法ではない. ベイズ学習の枠組みでは, 事前分布の超パラメータはドメイン知識を反映した上で固定値として設定するものである. 超パラメータの値もデータから学習したい場合, 超パラメータの対する事前分布を用意すれば完全なベイズの枠組みとして学習させることもできる.

未観測の値$x_\ast\in{0,1}$に対する予測分布を計算する. 後に事後分布を使った予測分布を求めるが, 便宜上, まずは事前分布$p(\mu)$を使った予測分布を計算する.

$$\begin{aligned} p(x_\ast)&=\int p(x_\ast\vert\mu)p(\mu)\text{d}\mu \ &=\int \text{Bern}(x_\ast\vert\mu)\text{Beta}(\mu\vert a,b)\text{d}\mu \ &=C_B(a,b)\int\mu^{x_\ast}(1-\mu)^{1-x_\ast}\mu^{a-1}(1-\mu)^{b-1}\text{d}\mu\ &=C_B(a,b)\int\mu^{x_\ast+a-1}(1-\mu)^{1-x_\ast+b-1}\text{d}\mu \ &=\frac{C_B(a,b)}{C_B(x_\ast+a,1-x_\ast+b)}\ &=\frac{\Gamma(a+b)\Gamma(x_\ast+a)\Gamma(1-x_\ast+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+1)} \end{aligned}$$

ベータ分布の正規化項 $C_B(a,b)\coloneqq \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$