README

PRML の第 1 章(とそれ以前の部分)について,

- これまであまり触れなかった事項
- ご確認・ご相談したい事項
- その他の重要事項

を簡単にまとめる.

以下，本書に対応した章立てになっている.

Mathematical notation

> The notation $g(x) = O(f(x))$ denotes that $|f(x)/g(x)|$ is bounded as $x \to \infty$

たぶん定義が間違ってて，$|f(x)/g(x)|$ じゃなくて $|g(x)/f(x)|$ だと思う.

Contents

ESLの目次と比較しても分かるように，ブースティング,ランダムフォレスト, DNN などの ML 定番の個々の手法は扱われてなかったりする．なのでそこは他を参照する必要がある．本書では，より本質的な基礎の部分に重点が置かれている.

1.1. Example: Polynomial Curve Fitting

ここの多項式回帰の例，機械学習・統計モデリングの重要な概念 (小西本でやった) を理解するのに打ってつけ．分かりやすい.

> By adopting a Bayesian approach, the over-fitting problem can be avoided. (P9)

ずっと気になっていた「ベイズは原理的に過適合しない」という話.
後の 3.4 節で詳しく理解したいが，雑に言うと「ベイズアプローチではそもそもデータにフィットさせようとせず，データを使って事前分布を更新するだけ」ということ．最尤法みたいに完全にデータに合わせにいくのではなく，事前に持っている情報をデータで更新する感じ.

1.2. Probability Theory

1.2.1. Probability densities

1.2.2. Expectations and covariances

この２つは青本とかでがっつり扱っている.

1.2.3 Bayesian probabilities

- 頻度主義的な確率：試行(観測)を無限回繰り返した時の割合
  - コインの表が出る確率
  - 母集団分布
- ベイズ的な確率：不確実性(可能性,適切さ)を 0〜1 で定量化したもの
  - コロナが5月中に収束する確率
  - パラメータの事前分布

> $\text{posterior} \propto \text{likelihood} \times \text{prior} \qquad (1.44)$

データ $\mathcal{D}$ からパラメータ $\boldsymbol{w}$ の事後分布を導出する場合だと，(1.43) より

$$p(\boldsymbol{w}|\mathcal{D}) \ \propto \ p(\mathcal{D}|\boldsymbol{w})\, p(\boldsymbol{w})$$

という式になる．後の MAP 推定を意識して解釈すると...

- 事前確率 $p(\boldsymbol{w})$ が高めな $\boldsymbol{w}$ は事後確率も大きくなりやすい.
  つまり，情報が与えられる(他の変数＝データが観測される)前から有力だと思われていた $\boldsymbol{w}$ 値は，情報が与えられた後も有力であり続ける.

- 尤度 $p(\boldsymbol{w})$ が高めな $\boldsymbol{w}$ は事後確率が大きくなりやすい.
  つまり，観測されたデータを発生させやすい(観測結果に対して辻褄が合うような) $\boldsymbol{w}$ 値は，情報が与えられた後には有力候補となる.
- まとめると，データが観測される前から有力候補で，かつ，観測されたデータと辻褄が合うような $\boldsymbol{w}$ 値が有力候補となる(事後確率の値が高くなる).

↓

このようにベイズでは「データ観測前に持っている主観」も考慮された上でパラメータが推定される.

↓

これは，例えば

> for instance, that a fair-looking coin is tossed three times and lands heads each time. A classical maximum likelihood estimate of the probability of landing heads would give 1, implying that all future tosses will land heads! By contrast, a Bayesian approach with any reasonable prior will lead to a much less extreme conclusion. (P23)

のような例を考えると，主観(=事前知識)を自然に用いることができるベイズアプローチは良さげ.

↓

ただ，主観によって得られる結果が変わり，極端に変な事前分布を設定してしまうと的外れな結論になることもある．これはベイズアプローチのデメリット.

↓

訳注の

> 頻度主義とベイズ主義の論争のポイントを一言で言えば「どこまで主観性を認めるか」という哲学的な問題となる.

が本質だと思う．先ほどのコインの例だと

- ベイズ主義：「公平に見えるんだし，それを事前分布として取り入れて極端な結論を避けようよ」
- 頻度主義：「いやいや公平に見えるなんていう主観(ある意味偏見)は取り除いて，純粋にデータだけを信じようよ」

というイメージ.

> In both the Bayesian and frequentist paradigms, the likelihood function p(D|w) plays a central role. However, the manner in which it is used is fundamentally dif- ferent in the two approaches. (P22)

- 頻度主義：パラメータの最尤推定の際に用いる.
- ベイズ：観測されたデータの情報からパラメータの分布を更新(事前→事後)する時に用いる.

1.2.4. The Gaussian distribution

> This might seem like a strange criterion because, from our fore- going discussion of probability theory, it would seem more natural to maximize the probability of the parameters given the data, not the probability of the data given the parameters. (P26)

ベイジアンぽい主張.

尤度すなわち「あるパラメータのもとで観測データが発生する確率」を見てパラメータ推定するのはなんか奇妙だよね，だったら事後分布すなわち「観測データのもとでのパラメータの分布」を見てパラメータ推定するほうが自然だよね，という感じ.

言われてみればそんな気もしてきた.

> particular, we shall show that the maximum likelihood approach systematically underestimates the variance of the distribution. This is an example of a phenomenon called bias and is related to the problem of over-fitting encountered in the context of polynomial curve fitting. (P27)

「正規分布の分散の最尤推定量にバイアスがあること」が「多項式回帰を最小二乗(=最尤)推定した際に over-fitting が発生しうること」は関連があるらしい．over-fitting は最尤法に起因するもの，みたいに書かれている．こういう感覚なかったので今後に理解.

1.2.5. Curve fitting re-visited

> $$p(\boldsymbol{w}|\mathbf{x},\mathbf{t},\alpha,\beta) \;\propto\; p(\mathbf{t}|\mathbf{x},\beta,\boldsymbol{w})\,p(\boldsymbol{w}|\alpha) \qquad (1.66)$$

これ，自明ではない気がしたので導出してみる.

$$\begin{aligned}
p(\boldsymbol{w}|\mathbf{x},\mathbf{t},\alpha,\beta) &\propto p(\mathbf{x},\mathbf{t},\alpha,\beta|\boldsymbol{w})\,p(\boldsymbol{w}) \\
&= p(\mathbf{t}|\mathbf{x},\alpha,\beta,\boldsymbol{w})\,p(\mathbf{x},\alpha,\beta|\boldsymbol{w})\,p(\boldsymbol{w}) \\
&= p(\mathbf{t}|\mathbf{x},\alpha,\beta,\boldsymbol{w})\,p(\mathbf{x}|\boldsymbol{w})\,p(\alpha|\boldsymbol{w})\,p(\beta|\boldsymbol{w})\,p(\boldsymbol{w}) \\
&= p(\mathbf{t}|\mathbf{x},\alpha,\beta,\boldsymbol{w})\,p(\mathbf{x})\,p(\alpha|\boldsymbol{w})\,p(\beta)\,p(\boldsymbol{w}) \\
&\propto p(\mathbf{t}|\mathbf{x},\alpha,\beta,\boldsymbol{w})\,p(\alpha|\boldsymbol{w})\,p(\boldsymbol{w}) \\
&= p(\mathbf{t}|\mathbf{x},\alpha,\beta,\boldsymbol{w})\,p(\alpha)\,p(\boldsymbol{w}|\alpha) \\
&= p(\mathbf{t}|\mathbf{x},\beta,\boldsymbol{w})\,p(\alpha)\,p(\boldsymbol{w}|\alpha) \\
&\propto p(\mathbf{t}|\mathbf{x},\beta,\boldsymbol{w})\,p(\boldsymbol{w}|\alpha)
\end{aligned}$$

各変形($=, \propto$)の意味を上から順に説明すると...

1. (1.43), (1.44) みたいなベイズの定理より．分母は $\boldsymbol{w}$ に依存しないので比例定数扱いでスルー.
2. 乗法定理(1.11)の少し拡張($\boldsymbol{w}$ を残してるから)．青本に載ってたと思う.
3. 説明変数データ $\mathbf{x}$，係数 $\boldsymbol{w}$ のブレ(事前分布の分散の逆数) α，誤差分散(の逆数) β は，直感的に考えて独立(と仮定しているんだと思う).
4. $\mathbf{x}, \boldsymbol{w}$ は独立で，$\boldsymbol{\beta}, \boldsymbol{w}$ も独立(と仮定しているんだと思う)
5. $\boldsymbol{w}$ の関数でない部分は比例定数扱いしてスルー.
6. ベイズの定理

7. $\boldsymbol{w}$ が与えられた下では $\mathbf{t}, \alpha$ は独立だと思う．事前分布の逆分散 α は $\boldsymbol{w}$ を経由して $\mathbf{t}$ に影響するが，その経由役である $\boldsymbol{w}$ がすでに与えられているので．
8. $\boldsymbol{w}$ の関数でない部分は比例定数扱いしてスルー．

結構めんどかったし，ここまで考えないと (1.66) って導けないのだろうか．
何か勘違いしていて，もっとシンプルな方法で導けるのだろうか．
全体的に「何と何を独立と仮定するか」が明示されていなくて少し困ったが，何らかの理由から自明なのだろうか．

> We can now determine $\boldsymbol{w}$ by finding the most probable value of $\boldsymbol{w}$ given the data, in other words by maximizing the posterior distribution. This technique is called maximum posterior, or simply MAP. (P30)

MAP推定は自然な発想．

頻度論ではデータからパラメータの値を点推定するが，ベイズではデータからパラメータの事後分布を導出する．
→ でもやっぱ，結局パラメータの値はいくらなの？という点推定はしたくなる．
→ そこで，事後分布で密度最大となるような点を推定値としよう．

MAP推定するためには (1.66) 右辺を最大化すれば良く，対数をとった上で変形していくと...

$$
\begin{aligned}
&\log[p(\mathbf{t}|\mathbf{x}, \beta, \boldsymbol{w})\, p(\boldsymbol{w}|\alpha)] \\
&= \log p(\mathbf{t}|\mathbf{x}, \beta, \boldsymbol{w}) + \log p(\boldsymbol{w}|\alpha) \\
&\propto -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \log p(\boldsymbol{w}|\alpha) \\
&= -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \log\left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} \\
&\propto -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}
\end{aligned}
$$

となる．途中で(1.62)と(1.65)を使っている．
さらに，マイナスをかけてハイパーパラメータ β を $\lambda = \frac{\alpha}{\beta}$ に置き換えて(β の代わりにこの λ を調整すると考えて)変形すると...

$$
\begin{aligned}
&\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} \\
&= \frac{\alpha}{2\lambda}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} \\
&= \frac{\alpha}{\lambda}\left(\frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}\right) \\
&\propto \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}
\end{aligned}
$$

となり，正則化回帰(リッジ回帰)の目的関数(1.4)式が導かれる.
つまり，今回の設定の下でのMAP推定(事後分布の最大化)は，頻度論の正則化回帰(罰則付き二乗誤差の最大化)と等価である.

これは直感的にも納得できる．なぜなら,

- 正則化：回帰係数パラメータの値が大きくなりすぎないように罰則を加えて，そのうえでパラメータ推定
- ベイズ：回帰係数パラメータは極端に大きい値は取りづらいような正規分布に従いそうでしょ，という事前情報(分布)を持っておいて，それに尤度(観測データの情報)による更新をかける

という感じで，結局やってることが近いから.

1.2.6. Bayesian curve fitting

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|\boldsymbol{w}, x)\, p(\boldsymbol{w}|\mathbf{x}, \mathbf{t})\, \mathrm{d}\boldsymbol{w} \qquad (1.68)$$

この予測事後分布(主観を含むモデルと観測データから導出された予測値の分布)の形をみると，ベイズによる予測はモデルアベレージングあるいはアンサンブル学習としても解釈できそう． $\boldsymbol{w}$ 値に応じて無数のモデル $p(t|\boldsymbol{w}, x)$ が考えられ，それらを $\boldsymbol{w}$ 値の確からしさ $p(\boldsymbol{w}|\mathbf{x}, \mathbf{t})$ で重み付けして足し合わせているイメージ.

1.3. Model Selection

だいたい知ってる話.

1.4. The Curse of Dimensionality

だいたい知ってる話.

1.5. Decision Theory

> Here we turn to a discussion of decision theory that, when combined with probability theory, allows us to make optimal decisions in situations involving uncertainty such as those encountered in pattern recognition. (P38)

モデル(分布)の推定はもう終わって，それを使って予測して意思決定を行おうとするフェーズのための議論.

小西本でロジスティック判別の閾値を決める時に扱われていた話だが，あれは決定理論という枠組みの一部だったのか.

1.5.1. Minimizing the misclassification rate

> We need a rule that assigns each value of $\boldsymbol{x}$ to one of the available classes. Such a rule will divide the input space into regions $\mathcal{R}_k$ called decision regions, one for each class, such that all points in $\mathcal{R}_k$ are assigned to class C_k. The boundaries between decision regions are called decision boundaries or decision surfaces. (P39)

決定理論は「モデルをもとに意思決定を行うためのもの」と言ったが，より具体的には「モデル(分布)をもとに決定領域(決定境界)をどう定めるべきか」を議論するための理論.

結局,
(仮定した分布のもとでの) 誤判別率を最小にする意思決定(判別)を行うためには事後分布最大のクラスに分類すれば良い，ということ.

1.5.2. Minimizing the expected loss

ざっくりまとめると...

ベイズアプローチで得られる事後予測分布と決定理論を組み合わせることで,
誤判別のコストを考慮した決定規則(決定領域)を決めることができる.

1.5.3. The reject option

棄却オプションは「確信が薄い場合において分類する(結論を出す)ことを諦める」こと．これ知らなかったが，確かに応用上は普通に便利そう．例えば

> For example, in our hypothetical medical illustration, it may be appropriate to use an automatic system to classify those X-ray images for which there is little doubt as to the correct class, while leav- ing a human expert to classify the more ambiguous cases.

という感じで，確信の高い時は自動で判別し，確信が薄い場合は棄却オプションをとって判断を専門家(医師)に委ねる，みたいな．

> rejecting those inputs $\boldsymbol{x}$ for which the largest of the posterior probabilities $p(C_k|\boldsymbol{x})$ is less than or equal to θ

これは自然．こう定式化した時に，

- $\theta = 1$：全ての入力が棄却される．
- $\theta < \frac{1}{K}$：全ての入力が棄却されない(常に判別決定が行われる)

というのも自明．図1.26 がわかりやすい．

1.5.4. Inference and decision

> we can identify three distinct approaches to solving decision problems, all of which have been used in practical applications. (P43)

生成モデル (generative model)

入力(説明変数)とクラスの同時分布 $p(\boldsymbol{x}, C)$ のモデルを作って，それをデータから推定する．そこから事後予測分布(クラス事後確率) $p(C|\boldsymbol{x})$ を導出し，あとは決定理論の枠組みを使って状況に応じた意思決定(判別や棄却オプション)をすることができる．周辺化すれば $p(\boldsymbol{x})$ も求まるので，予測の際に(外れ値)が入力されたら「外挿っぽい！」と注意を出せる．

識別モデル (discriminative model)

入力(説明変数) $\boldsymbol{x}$ の分布は考えず，事後予測分布(クラス事後確率) $p(C|\boldsymbol{x})$ を直接モデル化してデータから推定する．
あとは生成モデル同様に決定理論の枠組みを使って状況に応じた分類を行うことができる．同時分布までは考えない分，生成モデルよりは少ないサンプルサイズで済む．

識別関数 (discriminant function)

そもそも確率や分布なんて考えない．$C = f(\boldsymbol{x})$ のように入力 $\boldsymbol{x}$ から直接クラスラベル C を 決定(予測)するような関数 f をデータから推定して，それを使って分類をおこなう．確率分布って何？って人でも問題なく使える．（だが確率を使わないので決定理論を適用できない！）

↓

この3つの例としては...

- 生成モデル

    - ナイーブベイズ (8.2.2 節で詳しく)
- 識別モデル
    - ロジスティック判別
    - 出力層にソフトマックス関数を使ったニューラルネットワーク
- 識別関数
    - SVM
    - ランダムフォレスト
    - フィッシャー線形判別