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

> [p43] we can identify three distinct approaches to solving decision problems, all of which have been used in practical applications.

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) $p(C|\boldsymbol{x})$ を導出して利用する．

決定理論をフル活用し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 (もしくはクラス構成比率から直接推定できている) 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．

現象全体をモデリングするイメージ．作ったモデルから $\boldsymbol{x}, C$ を**生成**してシミュレーションを行うことができる．

- ナイーブベイズ (8.2.2 節で詳しく)

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.

決定理論を使って状況に応じた分類を行うことができる (生成モデルほど柔軟なことはできないが).
同時分布までは考えないので生成モデルよりは少ないサンプルサイズで済む．目的によっては識別モデルの方がコスパが高い場合もある.

- ロジスティック判別モデル
- 出力層の活性化関数として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 (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.

必要とされる数理(統計)の知識は生成モデルとかに比べて少ないが，確率を使わないので決定理論を適用できない．イメージ的には，とにかくより良い決定領域を作ろうとするアプローチ.

- SVM
- ランダムフォレスト
- フィッシャー線形判別 (クラス分布にガウスを仮定してマハラノビス距離使って，などとしっかり想定すれば生成モデルになる)

Combining models

> [p45] For example, in our hypothetical medical diagnosis problem, we may have information available from, say, blood tests as well as X-ray im- ages. Rather than combine all of this heterogeneous information into one huge input space, it may be more effective to build one system to interpret the X- ray images and a different one to interpret the blood data. As long as each of the two models gives posterior probabilities for the classes, we can combine the outputs systematically using the rules of probability.

生成モデルの現実的なメリット.

X線画像データと血液データの両方を使ったモデルを作りたい場合，両者を同じ DB に保存したり同時にメモリに乗せて学習を回したりしようとすると，システムが複雑になるしリソースもかかる.

ここで生成モデルのアプローチを使うと「X線画像データだけを使った生成モデル $P(C|\boldsymbol{x}_I)$」と「血液データだけを使ったモデル $P(C|\boldsymbol{x}_B)$」を別々で作ったあと，両者を自然にアンサンブルできる.
具体的には，(1.84)の条件付き独立性を認めれば,

$$
\begin{aligned}
P(C|\boldsymbol{x}_I, \boldsymbol{x}_B) &\propto P(\boldsymbol{x}_I, \boldsymbol{x}_B|C)\, P(C) \\
&= P(\boldsymbol{x}_I|C)\, P(\boldsymbol{x}_B|C)\, P(C) \\
&= \frac{P(C|\boldsymbol{x}_I)\, P(\boldsymbol{x}_I)}{P(C)} \, \frac{P(C|\boldsymbol{x}_B)\, P(\boldsymbol{x}_B)}{P(C)} P(C) \\
&\propto \frac{P(C|\boldsymbol{x}_I)\, P(C|\boldsymbol{x}_B)}{P(C)}
\end{aligned}
$$

のように $P(C|\boldsymbol{x}_I, \boldsymbol{x}_B)$ を構成できる.

1.5.5. Loss functions for regression

流れとしては,

1. 回帰タスクでアウトカム事後予測分布 $P(t|\boldsymbol{x})$ がモデル化して求まっているとして，予測値としてどういう t の値を選ぶのが良いのだろうか.
2. 直感的には，条件付き期待値 $E_t[t|\boldsymbol{x}]$ (つまり $P(t|\boldsymbol{x})$ の期待値) を予測値として選べば良さそう.
3. この決定規則の妥当性は「期待二乗誤差の最小化」の観点から保証される.

という感じで，これは 1.5.1 の分類タスクでの

> 1. 分類タスクでクラス事後予測分布 $P(C|\boldsymbol{x})$ がモデル化して求まっているとして，どのクラスに分類するのが良いだろうか.
> 2. 直感的には，$P(C|\boldsymbol{x})$ を最大にするクラス C に分類すれば良さそう.
> 3. その決定規則の妥当性は「誤分類確率の最小化」の観点から保証される.

と全く同じ構成の話.

> [p47] we can identify three distinct approaches to solving regression problems given, in order of decreasing complexity, by:

分類モデルと同様に，回帰モデルも次の３つに分類できる.

(a) 生成モデル (generative model)

入力 $\boldsymbol{x}$ とアウトカム t の同時分布 $P(\boldsymbol{x}, t)$ をどうにかしてモデル化してパラメータ推定して，そこから $P(t|x)$ を求める．さらに期待値とって条件付き期待値 $E[t|\boldsymbol{x}]$ を得てそれを予測値とする.

シミュレーションでデータを無限に生成できるようになるし，幅広い意思決定を支援できる．棄却オプションとか外れ値アラートとか．その分データや計算リソース，数学スキルが必要.

- (こんな手法は現実的じゃなさそうだが) $p(\boldsymbol{x}, t)$ をカーネル密度推定してゲットしてそれを使う.

(b) 識別モデル (discriminative model)

条件付き分布 $P(t|x)$ を直接モデル化して推定して，そこから条件付き期待値 $E[t|\boldsymbol{x}]$ を得て，それを予測値とする.

シミュレーションでデータ生成したり外れ値チェックしたりする必要がなければ，$P(t|x)$ だけモデル化する識別モデルのがコスパ良い.

- 等分散ガウスノイズを仮定した線形回帰モデル (説明変数側の分布も考えていたら生成モデル)

(c) 識別関数 (discriminant function)

そもそも確率や分布なんて考えず $t = y(\boldsymbol{x})$ のように入力 $\boldsymbol{x}$ から直接アウトカム値 t を 決定(予測)するような関数 y をデータから推定して，それを使って予測をおこなう.

必要な数理的知識は少ないが，確率を使わないので決定理論をによる意思決定はできない.

- ニューラルネットワーク (頑張って分布考えた場合を除く)
- GBDT(勾配ブースティング決定木)
    - XGBoost
    - LightGBM

1.6. Information Theory

> [p48] The amount of information can be viewed as the ‘degree of surprise’ on learning the value of x. If we are told that a highly improbable event has just occurred, we will have received more information than if we were told that some very likely event has just occurred, and if we knew that the event was certain to happen we would receive no information.

この「珍しい値(事象)が観測された時より多くの情報量が得られる」という思想を基にして話が展開される．例えば「今日は晴れた」より「今日は雪が降った」の方が情報量多い，という思想.

後で出てくるが，情報量を「符号化する時に割り当てるべきビット数(符号長)」と捉えるのもわかりやすい.

「情報量」の定義として対数を使った $h(x) = -\log p(x)$ が採用されたのは,

- 情報量は非負値であるべき
- 情報量は観測される確率の単調減少関数であるべき
- 独立な事象が起きた際に得られる情報量は，それぞれの事象の情報量の単純な和になるべき

といった要件を満たすため.

> [p49] This important quantity is called the entropy of the random variable x.

(1.92) の情報量 $h(x) = -\log p(x)$ は事象(確率変数の観測値)に対して定義されているもので，(1.93) のエントロピー $H[x] = -\sum_x p(x) \log p(x)$ は確率変数に対して定義されている．「確率変数 x の値が観測された時に得られる情報量の期待値」という意味．

> [p50] The noiseless coding theorem (Shannon, 1948) states that the entropy is a lower bound on the number of bits needed to transmit the state of a random variable.

これをエントロピーの解釈の 1 つとして捉えるのも良い．「なるべく効率的に符号化した時に必要となる平均符号長」という感じ．

> [p51] Distributions $p(x_i)$ that are sharply peaked around a few values will have a relatively low entropy, whereas those that are spread more evenly across many values will have higher entropy, as illustrated in Figure 1.30.

これもエントロピーの解釈として分かりやすい．例えば「誕生月」よりも「誕生日」の方がエントロピー(観測された時に得られる平均的な情報量)が大きい，となる．直感的．また「離散型確率変数において，とりうる値が固定されていれば，エントロピー最大となるのは一様分布」という結果も出ていて，自然．

> [p52] We can extend the definition of entropy to include distributions $p(x)$ over continuous variables x as follows.

ここ以降の流れをまとめる．

1. 離散型確率変数に定義したエントロピー (1.93) を連続型確率変数にも拡張したい
2. 和を量子化して積分に持っていこうとすると (1.103) の右辺第2項に発散する項 $-\ln \Delta$ が出てきてしまう．これは「連続値を厳密に記録しようとすると無限のビット数が必要となる」ことが反映されている．

   > This reflects the fact that to specify a continuous variable very precisely requires a large number of bits.

3. この発散項は分布の形状に関係ないので無視してしまって，第1項の方だけを見ることにし，一応名前を変えて「微分エントロピー」として (1.103) のように定義しよう．
4. 結局のところ，離散型のエントロピーの $\sum$ が自然に $\int$ に拡張されたことになった．

> [p53] In the case of discrete distributions, we saw that the maximum entropy con- figuration corresponded to an equal distribution of probabilities across the possible states of the variable. Let us now consider the maximum entropy configuration for a continuous variable.

ここ以降の流れをまとめる．

1. 離散型のときにエントロピーを最大にするのは，(とりうる値の数が同じなら) 一様分布だった．

では，連続型分布についてはどうだろう？
2. 条件を揃えるため「平均と分散同じ分布の中で」微分エントロピーを最大にするものを探すことにしよう．(1.106, 1.107 式)
3. ラグランジュ乗数法と変分法を使って計算すると，微分エントロピーを最大にするのはガウス分布だということがわかる．(一様分布でなく！) (離散型のエントロピーをそっくりそのまま拡張できたわけではないことが起因している？)

(1.110) の結果から

> [p54] Thus we see again that the entropy increases as the distribution becomes broader, i.e., as σ2 increases.

と言え，これもエントロピーの概念を押さえる上で直感的.

1.6.1. Relative entropy and mutual information

KL ダイバージェンスについて,

> [p57] Thus we can interpret the Kullback-Leibler divergence as a measure of the dissimilarity of the two distributions $p(x)$ and $q(x)$.

すなわち「分布間の隔たりの尺度」としての解釈と,

> [p57] If we use a distribution that is different from the true one, then we must necessarily have a less efficient coding, and on average the additional information that must be transmitted is (at least) equal to the Kullback-Leibler divergence between the two distributions.

すなわち「符号化の際に追加で必要な情報量」としての解釈の２つ述べられている.

それぞれについて，簡単にまとめる.

「分布間の隔たりの尺度」としての KL ダイバージェンス

これは直感的だが，定義式 (1.113) を少し変形して

$$\begin{aligned} KL(p||q) &= \int \{-\ln q(x) - (-\ln p(x))\}\ p(x)\, dx \\ &= E_p[-\ln q(X) - (-\ln p(X))] \end{aligned}$$

とすると，より分かりやすい．(ノーテーションは雑)

$-\ln q(x) - (-\ln p(x))$ は実現値 x における真の分布 p とモデル分布 q の隔たり(乖離度)みたいな意味．さらに，観測されやすい x における乖離度を重視したい，という気持ちで，期待値 E_p をとってる.

また，↓の KL ダイバージェンスの性質からも「隔たり」という感じがして良い．

- $KL(p||q) \geq 0$
- 等号成立条件は $p(x) = q(x)$

また，よくある注意点として，

> [p55] Note that it is not a symmetrical quantity, that is to say $KL(p||q) \neq KL(q||p)$.

というのがあるが，これはそんなに不自然ではない．真の分布で期待値をとる(すなわち真の分布で見て発生確率が高い場所での隔たりを重視する)ので，$p()$ と $q()$ のどちらを真の分布と見てどちらをモデル分布と見るかによって隔たりの値が変わる．これはそうであって欲しいし，自然．

「符号化の際に追加で必要な情報量」としての KL ダイバージェンス

こっちの解釈はちょっと理解足りていないが，イメージでまとめる．

1. 真の分布 $p(x)$ が分かっていれば，最小平均符号長 (＝エントロピー＝ $-\sum p(x) \ln p(x)$) を達成する符号化ルールを作れる．作り方としては，ある実現値 x に長さ $-\ln p(x)$ の符号を割り当てるイメージ．
2. ただ実際には，$p(x)$ は未知なので，モデル $q(x)$ で近似して，これをもとに符号化するしかない．ある実現値 x に長さ $-\ln q(x)$ の符号を割り当てる，という符号化を行うとすると，平均符号長は $-\sum p(x) \ln q(x)$ となる．
3. つまり結局，未知の真の分布 $p(x)$ をモデル $q(x)$ で近似して符号化してしまったせいで，追加で $-\sum p(x) \ln q(x) - (-\sum p(x) \ln p(x))$ の平均符号長を使うハメになってしまっている．このことを指して↓のような記述があるんだと思う．（ちなみに和訳ではここの "specify" を "特定" と訳しているが，"明示" とか "表現" というニュアンスのが近い気がする．）

> [p55] If we use $q(x)$ to construct a coding scheme for the purpose of transmitting values of x to a receiver, then the average additional amount of information (in nats) required to specify the value of x (assuming we choose an efficient coding scheme) as a result of using q(x) instead of the true distribution $p(x)$ is given by KL divergence (1.113)

次に，KL ダイバージェンスと最尤法の関係について見ていく．(1.119) の近似を少し変形すると，

$$\begin{aligned} KL(p||q) &= E_{X\sim p}[-\ln q(X|\theta) - (-\ln p(X))] \\ &\approx \frac{1}{N}\sum_{n=1}^{N}[-\ln q(x_n|\theta) - (-\ln p(x_n))] \end{aligned}$$

となるが，ここで近似 KL ダイバージェンスを最小にする θ を決める上では第2項は関係ないのでそれを無視して

$$\begin{aligned} KL(p||q) &\propto E_{X\sim p}[-\ln q(X|\theta)] \\ &\approx \frac{1}{N}\sum_{n=1}^{N}[-\ln q(x_n|\theta)] \end{aligned}$$

の部分を考えると，確かに「近似 KL ダイバージェンスの最小化は対数尤度の最大化と等価」てのは言える.

けれどもあくまで「KL ダイバージェンスの近似が妥当なとき」に限られる．やってることは期待値を標本平均で推定しているような感じなので，大体の場合問題ないが，**i.i.d. が崩れてしまうと**，色々とやばくなってしまう.

次に，相互情報量について．発想は自然で，流れとしては...

1. x, y が独立であることと $p(x,y) = p(x)p(y)$ は同値 (というか定義)
2. よって「x, y がどのくらい独立に近い(遠い)か」は「$p(x,y)$ に対する $p(x)p(y)$の近似精度」つまり「$p(x,y)$ と $p(x)p(y)$ の隔たり」で表せそう.
3. これは $p(x,y)$ と $p(x)p(y)$の KL ダイバージェンス $KL(p(x,y)||p(x)p(y))$ で定量的に測れるね.
4. これを x, y の相互情報量 $I(x,y)$ と呼んで「独立からの遠さ」の指標として使おう.