1 Jensen's inequality

$f(x)$ を convex downward とする． $\forall \lambda_n \geq 0, x_n (n = 1, ..., N) | \sum_{n=1}^{N} \lambda_n = 1$ のとき，

$$\sum_{n=1}^{N} \lambda_n f(x_n) \geq f(\sum_{n=1}^{N} \lambda_n x_n) \tag{1.1}$$

が成り立つ．これをを Jensen's inequality という．

Proof. convex downward の定義より，

$$\begin{aligned} &\forall \lambda, x_1, x_2 | 0 \leq \lambda \leq 1 \Rightarrow \\ &\lambda f(x_1) + (1 - \lambda) f(x_2) \geq f(\lambda x_1 + (1 - \lambda) x_2) \end{aligned} \tag{1.2}$$

なので，N=2 のときには Jensen's ineqality は明らかに成り立つ．
また，N=k のときに Jensen's inequality が成り立つとすると，

$$\begin{aligned} \sum_{n=1}^{k+1} \lambda_n f(x_n) &= \sum_{n+1}^{k} \lambda_n f(x_n) + \lambda_{k+1} f(x_{k+1}) \\ &\quad (\ A = \sum_{n=1}^{k} \lambda_n \text{として},) \\ &= A \sum_{n=1}^{k} \frac{\lambda_n}{A} f(x_n) + \lambda_{k+1} f(x_{k+1}) \\ &\geq A \sum_{n=1}^{k} \frac{\lambda_n}{A} f(x_n) + \lambda_{k+1} f(x_{k+1}) \\ &\geq A f(\sum_{n=1}^{k} \frac{\lambda_n}{A} x_n) + \lambda_{k+1} f(x_{k+1}) \quad (\because \sum_{n=1}^{k} \frac{\lambda_n}{A} = 1, \text{上の仮定}) \\ &\geq f(\sum_{n=1}^{k+1} \lambda_n x_n) \quad (\because A + \lambda_k = 1, \text{凸関数の定義}) \end{aligned} \tag{1.3}$$

よって帰納的に，任意の整数 $N > 2$ について Jnensen's inequality が成り立つ．Q.E.D.

□

直感的には，任意の凸関数上の点をとったときに，その点を結んでできる多角形の内部の点は凸関数上の点よりも上にあるということで理解できる．

2 Laplace's method

Laplace's mothod (smoothing) とは，ある分布 $q(\boldsymbol{\theta})$ を Gaussian で近似する手法である．ここでは，Laplace's mothod による分布の積分の近似を示す．[2]

パラメータ $\boldsymbol{\theta}$ の分布に関する積分を考える（n は標本数）

$$\int \exp(nq(\boldsymbol{\theta})) d\boldsymbol{\theta} \tag{2.1}$$

を考える．$\boldsymbol{\theta}$ の mode を $\hat{\boldsymbol{\theta}}$ とし，その近傍でテイラー展開する．(n が十分に大きければ $\boldsymbol{\theta}$ は mode に集中すると仮定して近似している)．$\left.\frac{\partial q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}} = 0$ であるので，

$$q(\boldsymbol{\theta}) = q(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T J(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + o((\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2) \tag{2.2}$$

-

ここで，

$$J(\hat{\boldsymbol{\theta}}) = -\left.\frac{\partial^2 q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\hat{\boldsymbol{\theta}}} \tag{2.3}$$

とした．

$$\int \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T J(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} = (2\pi)^{p/2} n^{-p/2} |J(\hat{\boldsymbol{\theta}})|^{-1/2} \tag{2.4}$$

より，結局，

$$\int \exp(nq(\boldsymbol{\theta}))d\boldsymbol{\theta} \approx (2\pi)^{p/2} n^{-p/2} |J(\hat{\boldsymbol{\theta}})|^{-1/2} \exp(nq(\hat{\boldsymbol{\theta}})) \tag{2.5}$$

と近似できる．

3 BIC

モデルの事後分布に基づくモデルの評価基準である BIC(Baysian Information criterion) を示す．一般に尤度を用いてモデルの比較を行うことはできない．なぜなら，尤度にはパラメータの次元などによるバイアスが含まれているからである [2]．モデルを複雑にすればするほど尤度を大きくすることができる問題からもこのことは明らかである．BIC では事後分布を考えることによりこの問題を解決してる．ただ，BIC では簡単にモデルの評価を行うことができるが，パラメータ事前分布が広がっており，そのヘッセ行列が非退化であるという仮定が多くの場合に妥当でないという問題もある [1].

P 次元のパラメータ $\boldsymbol{\theta}$ を持つモデル M のもとで観測データ $X = \{\boldsymbol{x_1}, ..., \boldsymbol{x_N}\}$ が観測されたときを考える．このとき，モデルの事後分布はベイズの定理より，

$$p(M|X) = \frac{p(X|M)p(M)}{\int p(X|M)p(M)dM} \tag{3.1}$$

となる．この事後確率を最大化するようなモデルを選ぶことが BIC の目標である．ここで，p(M) は一定とすると，分母は定数であるので，$p(X|M)$ を最大化すれば良いことがわかる．これをパラメータ $\boldsymbol{\theta}$ を用いてあわらすと，

$$p(X|M) = \int p(X|\boldsymbol{\theta})p(\boldsymbol{\theta}|M)d\boldsymbol{\theta} \tag{3.2}$$

となる．この周辺分布 $p(X|M)$ を Laplace's method(section2) 用いて近似する．

$L(\boldsymbol{\theta}) = \ln p(X|\boldsymbol{\theta}), \pi(\boldsymbol{\theta} = p(\boldsymbol{\theta}))$ として，ラプラス近似を行う．

$$
\begin{aligned}
p(X|M) &= \int \exp(L(\boldsymbol{\theta}))\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&\approx \int \exp\left\{L(\hat{\boldsymbol{\theta}}) - \frac{N}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T J(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\} \times \left\{\pi(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T \frac{\partial \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\} d\boldsymbol{\theta} \\
&\approx \exp(L(\hat{\boldsymbol{\theta}}))\pi(\hat{\boldsymbol{\theta}}) \int \exp\left\{-\frac{N}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T J(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} \\
&= \exp(L(\hat{\boldsymbol{\theta}}))\pi(\hat{\boldsymbol{\theta}})(2\pi)^{P/2} N^{-P/2} |J(\hat{\boldsymbol{\theta}})|^{-1/2}
\end{aligned} \tag{3.3}
$$

ただし,

$$
J(\hat{\boldsymbol{\theta}}) = -\frac{1}{N}\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \tag{3.4}
$$

として,

$$
\int (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \exp\left\{-\frac{N}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T J(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\} = 0 \tag{3.5}
$$

を用いた．対数を取って-2 を乗じて,

$$
-2\ln(p(X|M)) = -2L(\hat{\boldsymbol{\theta}}) - 2\ln(\pi(\hat{\boldsymbol{\theta}})) + P\ln N + \ln|J(\hat{\boldsymbol{\theta}})| \tag{3.6}
$$

N に関して $O(1)$ の項を無視して,

$$
BIC = -2L(\hat{\boldsymbol{\theta}}) + P\ln N \tag{3.7}
$$

を得る．

References

[1] Christopher Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. 1st ed. 2006. Corr. 2nd printing 2011. Springer, Feb. 2010. ISBN: 9780387310732. URL: http://amazon.co.jp/o/ASIN/0387310738/.

[2] 小西 貞則 and 北川 源四郎. **情報量規準** *(**シリーズ・予測と発見の科学**)*. 朝倉書店, Sept. 2004. ISBN: 9784254127829. URL: http://amazon.co.jp/o/ASIN/4254127820/.