黒木玄 **Gen Kuroki**
@genkuroki

mathtod.online/@genkuroki/2104...

ではKullback-Leibler情報量(もしくはダイバージェンス)に関するSanovの定理(大偏差原理のうちの一つ)は、大数の法則、中心極限定理の次くらいに来る必須の教養になるべきだという話をし、

mathtod.online/@genkuroki/2475...

では、正規分布で何らかの分布を近似するときに、誤差の尺度としてKLダイバージェンスを採用すると、誤差を最小にするためには、正規分布の期待値と分散を近似先の期待値と分散に一致させればよいことを説明した。正規分布

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

と期待値 $\mu_0$, 分散 $\sigma_0^2$ を持つ分布 $q(x)$ について、$\mu, \sigma$ を動かして、KLダイバージェンス

$$D(q||p) = \int q(x) \log \frac{q(x)}{p(x)}\, dx$$

を最小化するだけの簡単な計算です。

続く

2017年05月15日 13:15 · Web · 0 · 2 · Webで開く

---

黒木玄 **Gen Kuroki** @genkuroki on May 15
続き

その簡単な計算は指数型分布族に拡張されます。

正規分布は

$$p(x) = \frac{1}{Z} e^{-\beta x^2 - \gamma x}$$

と書き直せます。ここで、$\beta = 1/(2\sigma^2), \gamma = -\mu/\sigma^2$,

$$Z = \int e^{-\beta x^2 - \gamma x}\, dx.$$

これの一般化が指数型分布族です。

一般の指数型分布族は

$$
\begin{aligned}
p(x) &= \frac{1}{Z} e^{-\sum_{\nu=1}^{s} \beta_\nu f_\nu(x)} p_0(x), \\
Z &= \int e^{-\sum_{\nu=1}^{s} \beta_\nu f_\nu(x)} p_0(x)\,dx
\end{aligned}
$$

の形をしています。

分布族 $p(x)$ で何らかの分布 $q(x)$ をシミュレートするときボロが出る速さを最も遅くするためにはKLダイバージェンス

$$
D(q||p) = \int q(x) \log \frac{q(x)}{p(x)}\,dx
$$

を最小にすれば言い訳です(Sanovの定理)。

続く

---

黒木玄 **Gen Kuroki** @genkuroki on May 15

続き

指数型分布族は指数函数を使って書かれているので、対数と相性がよく、$p(x)$ を決めるパラメーター達 $\beta_\nu$ を動かして、$D(q||p)$ を最小にするためには

$$
\begin{aligned}
G &= -\int q(x) \log p(x)\,dx \\
&= \sum_\nu \beta_\nu \int f_\nu(x) q(x)\,dx - \log Z
\end{aligned}
$$

を最小にすればよいことがすぐにわかります。そして、

$$
\frac{\partial Z}{\partial \beta_\nu} = -\int f_\nu(x) p(x)\,dx
$$

なので、パラメーター達 $\beta_\nu$ を

$$
\int f_\nu(x) p(x)\,dx = \int f_\nu(x) q(x)\,dx
$$

が成立するように決めれば、$G$ が最小になり、したがって、$D(q||p)$ も最小になります。

続く

---

黒木玄 **Gen Kuroki** @genkuroki on May 15

続き

要するに、確率モデルとして指数型分布族 $p(x)$ を採用したケースでは、近似先の確率分布 $q(x)$ に関する函数 $f_\nu(x)$ の平均

$$
E_q[f_\nu(X)] = \int f_\nu(x) q(x)\,dx
$$

達と確率モデル $p(x)$ に関する函数 $f_\nu(x)$ の平均

$$E_p[f_\nu(X)] = \int f_\nu(x)p(x)\,dx$$

が一致するようにモデルのパラメーター達 $\beta_\nu$ を決めてやれば最良近似が得られるということです。

例：正規分布の場合には $f_1(x) = x$, $f_2(x) = x^2$ なので、確率モデルとして正規分布を採用することは「近似先の分布の1次と2次のモーメント以外の情報を忘れようとすること」に他ならないことになります。

これは指数型分布族の極めて特殊な性質です。だから、指数型分布族を例に何かについて(特にベイズ統計について)語る場合には厳しい注意が必要になります。

注意：統計力学では自動的に指数型分布族がカノニカル分布として出て来る。

---

黒木玄 **Gen Kuroki** @genkuroki on May 15

Kullback-Leiblerダイバージェンスに関する勉強において、指数型分布族 $p(x)$ と任意の分布 $q(x)$ についてKLダイバージェンス

$$D(q||p) = \int q(x)\log\frac{q(x)}{p(x)}\,dx$$

が最小になる条件を求めておくことは、算数で言えば一桁の足し算の練習をしているようなものだと思います。

そういうところでサボると色々わからなくなる。KLダイバージェンスが分布間の「距離」のようなものだという説明はよくあるのですが、計算例を何も知らないと感触をつかみようがないと思う。

証明を単に写経するのは時間の無駄。証明は理解しないと意味がない。証明を理解できないなら、がんばって計算例をたくさん作るのがよいと思う。手計算が無理ならコンピューターで数値計算しまくって納得するのもありだと思う。

私もいつもサボって色々わからなくなります。

---

黒木玄 **Gen Kuroki** @genkuroki on May 15

mathtod.online/@genkuroki/1135...

上のリンク先の訂正： $Z$ の対数微分になるべき部分が1ヶ所ただの微分になってしまっていました。正しくは

$$\frac{\partial\log Z}{\partial\beta_\nu} = -\int f_\nu(x)p(x)\,dx.$$

---

黒木玄 **Gen Kuroki** @genkuroki on May 15

表の出る確率が $p$ のコイン投げで表の出る確率が $q$ のコイン投げをシミュレートするときのKullback-Leiblerダイバージェンスは

$$D(q||p) = \\ q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$$

です。$n$ は大きいとする。表の出る確率が $p$ のコインを $n$ 回投げたときに、偶然、表の出る確率が $q$ のコインを投げたかのように見える確率の対数の $-1/n$ 倍がほぼ $D(q||p)$ になるというのがSanovの定理の特別な場合です。

極端な場合として、$p=0$ のコイン投げで $q=1/2$ のコイン投げをシミュレートする場合を考えましょう。表の出る確率が $0$ のコインをたくさん投げたとき、まるで表の出る確率が $1/2$ のコインを投げたかのように見える確率は $0$ です。実際、上の式に $q=1/2$ を代入し、$p\to 0$ とすると、$D(1/2||0)=\infty$ (「誤差」が無限大)になることがわかります。

続く

---

黒木玄 **Gen Kuroki** @genkuroki on May 15

続き

逆の極端な場合として、$p=1/2$ のコイン投げで $q=0$ のコイン投げをシミュレートする場合を考えましょう。表の出る確率が $1/2$ のコインを $n$ 回投げると、確率 $1/2^n$ で偶然すべてが裏になります。その対数の $-1/n$ 倍は $\log 2$ になります。上の式に $p=1/2$ を代入すると、

$$D(q||1/2) = q\log q + (1-q)\log(1-q) \\ + \log 2$$

なので $q\to 0$ とすることによって $D(0||1/2)$ も $\log 2$ になることがわかります。この $\log 2$ が確率 $1/2$ で表の出るコインで表が決して出ないコインをシミュレートしたときの「誤差」です。この場合には、確率の対数の $-1/n$ 倍とKLダイバージェンスがぴったり一致していますが、一般にはこうなりません。

大偏差原理では $n$ に関して指数函数的に減少する確率を扱っており、対数を取って $-1/n$ 倍することによって、確率が減少する速さを測っているわけです。

---

黒木玄 **Gen Kuroki** @genkuroki on May 15

大数の法則と中心極限定理だけでは足りない統計学の世界を理解するためには、Kullback-Leiblerダイバージェンスに関するSanovの定理について理解することが近道だと思う。

赤池情報量規準AICおよびその変種について理解するためには(AICが何を近似しているかを知っていれば当然のことなのですが)、KLダイバージェンスに関する理解が必須。

ベイズ統計学の基礎付けが書いてある渡辺澄夫『ベイズ統計の理論と方法』もKLダイバージェンスについて知らないと最初から何をやっているか理解不能になると思う。

そこで私が用意した易しい解説がこれ↓
github.com/genkuroki/Sanov