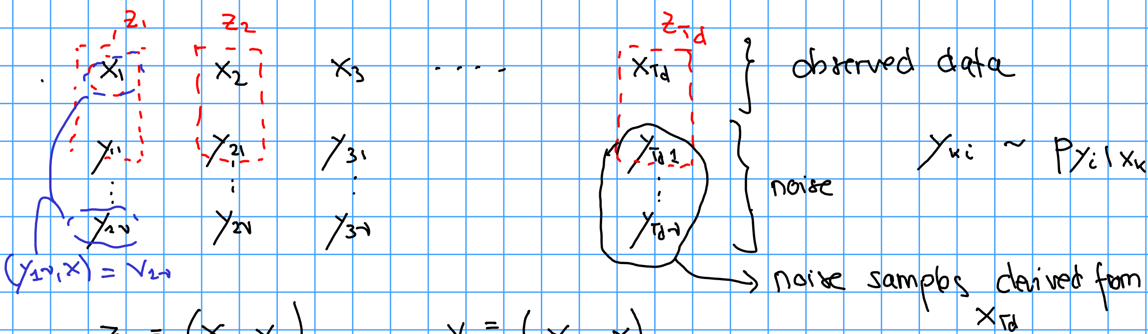
Better formulation of NCE with data-dependent noise distribution

Michael Gutmann   June 28 2012

---

given sample $(x_1 \dots x_n)$, generate ν noise samples for each data point:



$y_{ki} \sim p_{y_i|x_k} = p_{y|x}$ for simplicity

(one could take different distributions for different i and k!)

$$z = (x, y) \qquad v = (y, x)$$

$$p_z(u) = p_{y|x}(u_2|u_1)\, p_x(u_1) \quad ; \quad p_v(u) = p_{y|x}(u_1|u_2)\, p_x(u_2)$$

$$u = (u_1, u_2)$$

class one: observations $z_1 \dots z_{T_d} = (x_1, y_{11}) \dots (x_{T_d}, y_{T_d 1})$; $z_k \sim p_z$

class zero: observations $v_{1i} \dots v_{T_d i} = (y_{1i}, x_1) \dots (y_{T_d i}, x_{T_d})$ $\quad i = 2 \dots \nu$; $v_{ki} \sim p_v$

Class one has T_d members. Class zero has $(\nu - 1) \cdot T_d$ members.

$$P(C=1) = \frac{T_d}{(\nu-1)T_d + T_d} = \frac{T_d}{\nu T_d} = \frac{1}{\nu}$$

$$P(C=0) = \frac{(\nu-1)T_d}{\nu T_d} = 1 - \frac{1}{\nu} = \frac{\nu-1}{\nu}, \qquad P(C=0)/P(C=1) = (\nu-1)$$

$$p(u|C=1;\theta) = p_m(u_1;\theta)\, p_{y|x}(u_2|u_1) \qquad p(u|C=0;\theta) = p_m(u_2;\theta)\, p_{y_i|x}(u_1|u_2)$$

Posteriors:
$$P(C=1|u) = \frac{p(u|C=1)P(C=1)}{p(u|C=1)P(C=1) + p(u|C=0)P(C=0)}$$

$$= \frac{1}{1 + \frac{p_m(u_2;\theta)\, p_{y_i|x}(u_1|u_2)}{p_m(u_1;\theta)\, p_{y_1|x}(u_2|u_1)} \underbrace{\frac{P(C=0)}{P(C=1)}}_{=\nu-1}}$$

$$P(C=0|u) = 1 - P(C=1|u)$$

$$\left(1 - \frac{1}{1+x} = \frac{1+x}{1+x} - \frac{1}{1+x} = \frac{x}{1+x} = \frac{1}{1+1/x}\right)$$

$$= \frac{1}{1 + \frac{p_m(u_1;\theta)\, p_{y_1|x}(u_2|u_1)}{p_m(u_2;\theta)\, p_{y_i|x}(u_1|u_2)} \frac{1}{\nu-1}}$$

Logistic regression to distinguish between class 1 and class zero.

$$J_T(\theta) = \frac{1}{T_d}\sum_{t=1}^{T_d}\left[\ln P(C=1|z_t) + \sum_i \ln P(C=0|v_{it})\right]$$

$$= \frac{1}{T_d}\sum_{t=1}^{T_d} \ln P(C=1|z_t) + \frac{1}{T_d}\sum_{t=1}^{T_d}\sum_{i=2}^{\nu} \ln P(C=0|v_{it})$$

$$\hat{\theta}_T = \arg\max J_T(\theta)$$

Relation to NCE:

"Data": $(x_1, y_{11}) \ldots (x_{T_d}, y_{1T_d})$

"Noise": $(y_{1i}, x_1) \ldots (y_{T_d i}, x_{T_d})$ $\quad i = 2 \ldots \nu$

"Data" and "noise" are not independent. This means that the formula for the asymptotic error cannot be read at from the NCE theorems.

This relation to NCE suggests that $\hat{\theta}_T$ is consistent. On the next page, I consider the nonparametric case, which allows to prove consistency.

Shorthands:

$$f_m(\cdot;\theta) = \ln p_m(\cdot;\theta)$$

$$f_n(u;\theta) = \ln\left[p_{y|x}(u_1|u_2) / p_{y|x}(u_2|u_1)\right]$$

$$h(u;\theta) = 1 + \exp\left(f_m(u_2;\theta) - f_m(u_1;\theta) + f_n(u)\right)\cdot(\nu-1)$$

$$\check{h}(u;\theta) = 1 + \exp\left(f_m(u_1;\theta) - f_m(u_2;\theta) - f_n(u)\right)\frac{1}{\nu-1}$$

$$r(u;\theta) = \ln h(u;\theta)$$

$$\check{r}(u;\theta) = \ln \check{h}(u;\theta)$$

Case of large T_a :

$$\bar{J}(\theta) = \int \ln P(c=1|z)\, p_z(z)\, dz + \sum_{i=2}^{\nu} \int \ln P(c=0|v_i)\, p_v(v_i)\, dv_i$$

$$= \int \ln P(c=1|z)\, p_z(z)\, dz + (\nu-1) \int \ln P(c=0|v)\, p_v(v)\, dv$$

Reminder:

$z = (x, y) \qquad v = (y, x)$

$p_z(u) = p_{y|x}(u_2|u_1)\, p_x(u_1) \quad ; \quad p_v(u) = p_{y|x}(u_1|u_2)\, p_x(u_2)$

$$J(\theta) = -\int \ln\left[1 + \exp\left(f_m(u_2;\theta) - f_m(u_1;\theta) + f_n(u)\right)\cdot(\nu-1)\right] p_z(u)\, du$$

$$-(\nu-1)\int \ln\left[1 + \exp\left(f_m(u_1;\theta) - f_m(u_2;\theta) - f_n(u)\right)\frac{1}{\nu-1}\right] p_v(u)\, du$$

Computation of functional derivative

$\frac{1}{1+\frac{1}{a}\exp(-v_0)}$

- $\ln(1 + a\exp(v)) = \ln(1 + a\exp(v_0)) + \frac{a\exp(v_0)}{1+a\exp(v_0)}(v - v_0) + \frac{1}{1+{}^1/_a\exp(-v_0)}\,\frac{1}{1+a\exp(v_0)}\,\frac{1}{2}(v-v_0)^2$

$$\frac{d}{dv}\left(\frac{1}{1+\frac{1}{a}\exp(-v)}\right) = -\left(\frac{1}{1+\frac{1}{a}\exp(-v)}\right)^2 (-1)\,\frac{1}{a}\exp(-v)$$

$$= \frac{1}{1+\frac{1}{a}\exp(-v)}\;\frac{\frac{1}{a}\exp(-v)}{1+\frac{1}{a}\exp(-v)}$$

$$= \frac{1}{1+{}^1/_a\exp(-v)}\;\frac{1}{1+a\exp(v)}$$

- First integral in J : $v_0 = f_m(u_2) - f_m(u_1) + f_n(u)$

$v - v_0 = \varepsilon\left(\phi(u_2) - \phi(u_1)\right)$

$a = \nu - 1$

$1 + a\exp(v_0) = h(u)$

$\Rightarrow$

$1 + {}^1/_a\exp(-v_0) = \check{h}(u)$

• Second integral in J : $\quad v_0 = fm(u_1) - fm(u_2) - fn(u)$

$$v - v_0 = \varepsilon\left(\phi(u_1) - \phi(u_2)\right)$$

$$a = 1/(\nu-1)$$

$$\Rightarrow \quad \begin{aligned} 1 + a\exp(v_0) &= \check{h}(u) \\ 1 + 1/a \exp(-v_0) &= h(u) \end{aligned}$$

$$\cdot\ \tilde{J}(fm + \varepsilon\phi) = J(fm) + \int \frac{1}{\check{h}(u)}\ \varepsilon\left(\phi(u_1) - \phi(u_2)\right) p_Z(u)\,du$$

$$- \int \frac{(\nu-1)}{h(u)}\ \varepsilon\left(\phi(u_1) - \phi(u_2)\right) p_V(u)\,du$$

this terms are negative
$\Rightarrow$ negative curvature everywhere
$\Rightarrow$ critical point is a maximum.

$$\left\{ \begin{aligned} &- \int \frac{1}{\check{h}(u)\,h(u)}\ \frac{1}{2}\ \varepsilon^2\left(\phi(u_1) - \phi(u_2)\right)^2 p_Z(u)\,du \\ &- \int \frac{(\nu-1)}{\check{h}(u)\,h(u)}\ \frac{1}{2}\ \varepsilon^2\left(\phi(u_1) - \phi(u_2)\right)^2 p_V(u)\,dv \end{aligned} \right. + O(\varepsilon^3)$$

First order term is zero if : $\quad u_1 = u_2 \quad$ or $\quad u_1 \neq u_2$ and

$$\frac{1}{\check{h}(u)}\ p_Z(u) = \frac{(\nu-1)}{h(u)}\ p_V(u)$$

Reminder :

$$h(u;\theta) = 1 + \exp\left(fm(u_2;\theta) - fm(u_1;\theta) + fn(u) \right) \cdot (\nu-1)$$

$$\check{h}(u;\theta) = 1 + \exp\left(fm(u_1;\theta) - fm(u_2;\theta) - fn(u) \right) \frac{1}{\nu-1}$$

$$p_Z(u) = p_{Y|X}(u_2|u_1)\,p_X(u_1) \quad ; \quad p_V(u) = p_{Y|X}(u_1|u_2)\,p_X(u_2)$$

$$\Rightarrow \quad p_Z(u) = \frac{(\nu-1)\,\check{h}(u)}{h(u)}\ p_V(u) \qquad (1)$$

$$\frac{p_Z(u)}{p_V(u)} = \frac{(\nu-1) + \exp\left(fm(u_1) - fm(u_2) - fn(u) \right)}{1 + (\nu-1)\exp\left(fm(u_2) - fm(u_1) + fn(u) \right)}$$

$$\frac{p_X(u_1)\,p_{Y|X}(u_2|u_1)}{p_X(u_2)\,p_{Y|X}(u_1|u_2)} = \frac{(\nu-1) + \exp\left(fm(u_1) - fm(u_2) \right) \dfrac{p_{Y|X}(u_2|u_1)}{p_{Y|X}(u_1|u_2)}}{1 + (\nu-1)\exp\left(fm(u_2) - fm(u_1) \right) \dfrac{p_{Y|X}(u_1|u_2)}{p_{Y|X}(u_2|u_1)}}$$

sine

$$f_n(u;\theta) = \ln\left[p_{y|x}(u_1|u_2) \,/\, p_{y|x}(u_2|u_1) \right]$$

Note: For the above equation to make sense, you need $p_{y|x}(u_2|u_1) \neq 0$ and $p_{y|x}(u_1|u_2) \neq 0$. Else we find the trivial relation $0=0$ in Eq. (1)

We obtain:

$$\frac{p_x(u_1)}{p_x(u_2)} = \frac{(\nu-1)\,p_{y|x}(u_1|u_2) + \exp(\overbrace{f_{m1}}^{f_m(u_1)} - \overbrace{f_{m2}}^{f_m(u_2)})\, p_{y|x}(u_2|u_1)}{p_{y|x}(u_2|u_1) + (\nu-1)\exp(f_{m2}-f_{m1})\, p_{y|x}(u_1|u_2)}$$

$$= \exp(f_{m1}-f_{m2}) \left[\frac{(\nu-1)\,p_{y|x}(u_1|u_2)\exp(f_{m2}-f_{m1}) + p_{y|x}(u_2|u_1)}{p_{y|x}(u_2|u_1) + (\nu-1)\exp(f_{m2}-f_{m1})\, p_{y|x}(u_1|u_2)} \right]$$

$$= \frac{\exp(f_m(u_1))}{\exp(f_m(u_2))} \cdot 1$$

$$\Rightarrow \quad \frac{p_x(u_1)}{\exp(f_m(u_1))} = \frac{p_x(u_2)}{\exp(f_m(u_2))} \tag{2}$$

$$\Rightarrow \quad \frac{p_x(u_1)}{\exp(f_m(u_1))} = \text{constant}$$

$$\Rightarrow \quad \ln p_x = f_m + \text{constant}, \quad \text{or} \quad \boxed{f_m = \ln p_x + \text{constant}}$$

Note Eq (2) means: $\ln p_x(u_1) - \ln p_x(u_2) = f_m(u_1) - f_m(u_2) \quad \forall\, u_1 \neq u_2$, which can be considered a stronger version of score matching.

This proves that $\tilde{J}(f_m)$ attains its maximum at $\exp(f_m) \propto p_x$.