Как делать ансамблирование.

Есть произвольная байесовская дискриминативная модель.

$$p(y|x,w)p(w)$$

$$x^*, y^* - ?\ p(y^*, X, Y)$$

$$p(y^*, X, Y) = \int p(y^*|x^*, w)p(w|x,y)dw =$$

$$E_{p(w|x,y)}p(y^*|x^*, w) \text{ обычное правдоподобие}$$

Усреднять надо вероятности, и ни что иное.

Если есть MCMC, то оцениваем так:

$$E_{p(w|x,y)}p(y^*|x^*, w) \approx \frac{1}{K}\sum_k p(y^*|x^*, w_k)$$

Если есть только вар. вывод:

$$E_{p(w|x,y)}p(y^*|x^*, w) \approx E_{q(w)}p(y^*|x^*, w)$$

Основная тема – лок. репарам и почему она снижает дисперсию градента. Рассмотрим два случая:

1. есть елбо:

$$\mathcal{L} = E_{q(\dots)} \log p(y|x,w) = KL$$

2. некрасивое L:

$$L = E_{q(\dots)} \log p(y|x,w) \approx \frac{N}{M}\sum_{m=1}^{M} \log p(y_m|x_m, w); w \sim q(w); \frac{N}{M} \text{ – скалирование}$$

берем один семпл и используем для всех объектов минибатча

$$B = AW$$

$$\text{размерности: } MxD, MxD, DxD$$

При такой оценке появляется паразитная дисперсия и надо понять, как от нее избавиться. Есть $L_m = \log p(y_m|x_m, w)$. Допустим по индексам они одинаково распределены. $\hat{L} = \frac{N}{M}\sum_{m=1}^{M} L_m$.

Нужно получить дисперсию $\hat{L}$:

$$D\hat{L} =$$

$$D\left(\frac{N}{M}\sum_{m=1}^{M} L_m\right) =$$

$$\left(\frac{N}{M}\right)^2 \left(\sum_m D(L_m) + \sum_{i<j} \mathrm{cov}(L_i, L_j)\right) =$$

$$\frac{N^2}{M} D(L_M) + \frac{N^2}{M}(M-1)\mathrm{cov}L(L, L)$$

$$= O\left(\frac{1}{m}\right) + O(1)$$

Все "хреново"со вторым членом. Давайте это поправим. В лоб семплировать матрицу весов неэффективно и плохо.

У нас есть случайные веса $W \sim N(\mu, \sigma^2)$ и у нас есть $B$. Значит у нас есть нормальное распределение:

$$b_{mi} \sim N(b_{mi}|a_m^T\mu_i, a_m^T\sigma_i^2) =$$
$$N(b_{mi}|m, s^2) =$$
$$m = A\mu$$
$$s^2 = A^2\sigma^2$$
$$b = m + \varepsilon \odot s$$

Ковариация убита в ноль, все круто и быстро. Матрица весов больше нигде не участвует. Сначала получаем матрицы среднего и дисперсии и потом получаем семплы.

Можно считать, что $(W, b)$ – случайная величина. Мы интегрируем по $W$:

$$E_{q(w)}L_m = E_{q(w,b)}L_m = E_{q(b)}L_m$$

Понижает дисперсию, потому что больше нет дисперсии. Даже когда один вектор на вход подается.

here goes pic. 1

$$\text{RT:}$$
$$\varepsilon \sim p(\varepsilon) - \mathbb{R}^D$$
$$b = a^T\mu + a^T(\varepsilon \odot \sigma)$$
$$\text{LRT:}$$
$$\varepsilon \sim p(\varepsilon) - \mathbb{R}$$
$$b = a^T\mu + \varepsilon\sqrt{(a^2)^T\sigma^2}$$

Совсем другой результат. Процедура обучение неизменна. Все это только для понижения дисперсии. Даже в случае одного объекта.

here goes pic. 2

Рассмотрим линейную классификацию. Веса задают разделяющую границу. Семплируем новую матрицу весов. Она одинаковая для всех объектов. Это плохо, тк все градиенты будут скореллированны. Нам нужно не $\mu$, а все остальное. А если каждый будет тянуть в свою сторону, то в среднем они дадут нужный градиент.

Как у нас выглядит градиент лосса по вар параметрам? Будем считать, что случайность дальше не зависит от случайности по горизонтали и по вертикали.

$$\text{RT: } \frac{\partial L}{\partial \mu_i}, \frac{\partial L}{\partial \sigma_i^2}$$

$$\text{LRT: } \frac{\partial L}{\partial \mu_i}, ???$$

Первая производная такая же, а вторая самая интересная.

$$\text{RT: } \frac{\partial L}{\partial \sigma_i^2} = \frac{\partial L}{\partial b} \varepsilon_i q_i \frac{1}{2\sigma_i}$$

$$\text{LRT: } \frac{\partial L}{\partial \sigma_i^2} = \frac{\partial L}{\partial b} \varepsilon \frac{a_i^2}{2\sqrt{(a^2)^T \sigma^2}}$$

Law of total variance:

$$x, y; \quad D_y = D_x \left[E_{y|x} y\right] + E_x \left[D_{y|x} y\right]$$

Считаем дисперсию. Какие есть случайные величины? – $\varepsilon$+все остальное, $b$:

$$x, y; \quad D_{\varepsilon_{b_c}} L' = D_b \left[E_{\varepsilon, c|b} L'\right] + E_b \left[D_{\varepsilon, c|b} L'\right]$$

$$E_{e,c|b} L' = E_{c|b} \left(\frac{\partial L}{\partial b}\right) E_{\varepsilon|b}(\varepsilon_i) a_i \frac{1}{2\sigma_i}$$

$$E_{\varepsilon_i|b} \varepsilon_i = 0; \quad p(b, \varepsilon) = p(\varepsilon) p(b|\varepsilon) = N(0, 1)$$

$$\operatorname{cov}(\varepsilon_i, b) = \operatorname{cov}(\varepsilon_i, a^T \mu + a^T (\varepsilon \odot \sigma))$$

$$\begin{bmatrix} \varepsilon_i \\ b \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ a^T \mu \end{bmatrix}; \begin{bmatrix} 1 & a_i \sigma_i \\ (a_i \sigma_i)^T & \left(a^T\right)^2 \sigma^2 \end{bmatrix}\right)$$

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} a \\ b \end{bmatrix}; \begin{bmatrix} S_{xx} S_{xy} \\ S_{yx} S_{yy} \end{bmatrix}\right); \quad p(x|y) = N\left(x | a + \frac{S_{xy}(y - b)}{S_{yy}}, S_{xx} - \frac{S_{xy}^2}{S_{yy}}\right)$$

$$E_{\varepsilon_i|b} \varepsilon_i = \frac{a_i \sigma_i (b - a^T \mu)}{(a^T)^2 \sigma^2}$$

Запишем условную дисперсию:

$$D_{\varepsilon, c|b} L' = \left(E_{c|b} \left(\frac{\partial L}{\partial b}\right)\right)^2 D_{\varepsilon|b} \frac{\partial b}{\partial \sigma_i^2} + \left(E_{\varepsilon|b} \frac{\partial b}{\partial \sigma_i^2}\right)^2 D_{c|b} \frac{\partial L}{\partial b} +$$

$$\left(D_{\varepsilon|b} \frac{\partial b}{\partial \sigma_i^2}\right) \left(D_{c|b} \frac{\partial L}{\partial b}\right)$$

$$\begin{bmatrix}\varepsilon\\ b\end{bmatrix} \sim N\left(\begin{bmatrix}0\\ a^T\mu\end{bmatrix}; \begin{bmatrix}1 & \sqrt{(a^T)^2\sigma^2}\\ \sqrt{(a^T)^2\sigma^2} & (a^T)^2\sigma^2\end{bmatrix}\right)$$

$$\varepsilon|b \sim N\left(\varepsilon\middle|\frac{b-a^T\mu}{\sqrt{(a^T)^2\sigma^2}}, 0\right)$$

$$D^{\mathrm{LRT}}_{\varepsilon,c|b}L' = 0 + \left(E_{\varepsilon|b}\frac{\partial b}{\partial\sigma_i^2}\right)^2 D_{c|b}\frac{\partial L}{\partial b} + 0$$

Не совсем легально применять на свертках, но все все равно применяют. на бешках не отдельные нейроны, а целые картинки. мы семплируем для каждого окошка. и тогда домножать кл на размер окошка. но тогда все не работает. На практике: применяем conv2d на $a$, применяем conv2d на $a^2$. лол.

---

еще немного про лог-ю

$$q(w) = N(\mu, \alpha\mu^2)$$

$$p(w)\mathrm{KL}(q\|p) = -\frac{1}{2}\log\alpha\mu^2 - E_q\log p(w)$$

$$= -\frac{1}{2}\log\alpha\mu^2 - E_{\varepsilon\sim N(0,1)}\log p(\mu(1+\sqrt{a}\varepsilon)) - \frac{1}{2}\log\mu^2 - E_\varepsilon\log p(\mu(1+\sqrt{a}\varepsilon)) =$$

$$F(\alpha\mu) = [\alpha=0] = -\frac{1}{2}\log\mu^2 - \log p(\mu)\bigg|_{\frac{\partial}{\partial\mu}}$$

$$= -\frac{M}{\mu^2} - (\log p(\mu))' = 0$$

$$(\log p(\mu))' = -\frac{1}{\mu}; \quad \log p(w) = -\log|w| + C$$