p 30

(IV)

$$p(x|\eta) = h(x)\exp\left(\eta^T t(x) - a(\eta)\right)$$ Exponential family.

Dependent — natural param.

some suff. — sufficient stat (great) — log normalizer.

base measure.

<u>mean field variational model:</u> Approximate Posterior:

variational Objective function

$$v^* = \arg\min_v KL\left(q(\beta, z|v) \,\|\, p(\beta, z|x)\right)$$

(good read)

$$\mathcal{L}(v) = E\left[\log p(\beta, z, x|\eta)\right] - E\left[\log q(\beta, z|v)\right]$$

<u>mean field variational family:</u>

$$q(\beta, z|v) = q(\beta|\lambda)\prod_{n=1}^{N} q(z_n|\phi_n)$$

$$v = \{\lambda, \phi_{1-N}\}$$ easy.

<u>coordinate Ascent Varia. Inference:</u>

$$\lambda^* = E_q\left[\eta_g(z, x)\right]$$ global

$$\phi_n^* = E_q\left[\eta_L(\beta, x_n)\right]$$ local

$$q(\mu, z) = \prod_{i=1}^{k} q(\mu_k|\lambda_k)\prod_{n=1}^{N} q(z_n|\phi_n)$$

P 30

(v)

Model Criticism:

Exploration & prediction.

⇓

inference about hidden varg.

distribution

$$P(x_{new}|x) = \int \underbrace{p(\beta|x)}_{\text{not available}} \left(\int p(z_{new}|\beta)\, p(x_{new}|z_{new},\beta)\, dz \right) d$$

Predictive Sample Reuse:

n removed

$P(\beta, z, x)$

$$l_n = \log P(x_n | x_{[n]})$$

$P(\beta, z | x_{[n]})$

$$\left[= \log \int \left(\int p(x_n|z_n)\, q(z_n)\, dz \right) q_{[n]}(\beta)\, d\beta \right] q_{[n]}(\beta, z)$$

full likelihood $\sum_{n=1}^{N} l_n$

Posterior Predictive Check:

Test statistics

$$PPC = P(T(x^{rep}) > T(x) | x)$$

⇓

Data drawn from hypothetical feature. obs.

$$PPC = P(T(x^{rep}, \beta) > T(x, \beta) | x)$$

$$P(\beta, x^{rep} | x) = P(\beta|x)\, P(x^{rep}|\beta)$$

$$PPC = \int p(\beta|x) \int p(x^{rep}|\beta)\, \mathbb{1}_{[T(x^{rep},\beta) > T(x,\beta)]}\, dx^{rep}\, d\beta$$

P 10

$$T(x^{+}, \beta^{+}) > T(x ; \beta^{+})$$

$$T(x,\beta) = \frac{1}{N} \sum_{n=1}^{N} \log(x_n | \beta)$$ // may be T)

PPC is adaptive.

Basics

Batch Normalization in CNN: input $[N, C, H, W] \Rightarrow [i, c, j, k]$

N → Batch size; H → Height; W → width; C → filter channel

for all items; same filter (c)

$$\Rightarrow \mu_{B,c} = \frac{1}{NHW} \sum_{i=1}^{N} \sum_{j=1}^{H} \sum_{k=1}^{W} x_{i,c,j,k}$$

for each channel we get 1 value
So total C value.

$$\Rightarrow \sigma^2_{B,c} = \frac{1}{NHW} \sum_{i=1}^{N} \sum_{j=1}^{H} \sum_{k=1}^{W} \frac{(x_{i,c,j,k} - \mu_{B,c})^2}{\cancel{N} \; 1}$$

$$\Rightarrow \text{final output} \quad \hat{x}_{i,c,j,k} = \frac{x_{i,c,j,k} - \mu_{B,c}}{\sqrt{\sigma^2_{B,c} + \epsilon}}$$

⇒ Do calculation for each $c \in C$ channels.

More formally,

$$\hat{x}_{i,c,j,k} = \gamma \left(\frac{x_{i,c,j,k} - \mu_{B,c}}{\sqrt{\sigma^2_{B,c} + \epsilon}} \right) + \beta.$$

Solves:
(i) Internal Covariate Shift. (each ~~..~~ zero mean, var = 1)
→ features distribution differs internally.
→ inside the neural Network (layer - layer)
(ii) Robust Network creation → less prone to perturbation
(iii) Learning faster.

Basic

Instance Normalization (IN) / Layer Normalization

$$IN(x_{i,c,j,k}) = \gamma \left(\frac{x_{i,c,j,k} - \mu_{i,c}}{\sqrt{\sigma_{i,c}^2 + \epsilon}} \right) + \beta$$

→ can be conditioned → conditional IN

where,

$$\mu_{i,c} = \frac{1}{HW} \sum_{j=1}^{H} \sum_{k=1}^{W} x_{i,c,j,k}$$ → summed out.

$$\sigma_{i,c}^2 = \frac{1}{HW} \sum_{j=1}^{H} \sum_{k=1}^{W} \left(x_{i,c,j,k} - \mu_{i,c} \right)^2$$

Instance : for each n & c

Adaptive Instance Normalization (AdaIN)

→ var

Adaptive term.

$$AdaIN(x_{i,c}, y) = \sigma(y) \left(\frac{x_{i,c} - \mu_{i,c}}{\sqrt{\sigma_{i,c}^2 + \epsilon}} \right) + \mu(y)$$

mean → y

contional

→ 0 if y is constant → retain if y is constant

→ High if y varies a lot. → should be 0 mean.

Alternative

Layer Normalization.

$$LN(x_{i,c,j,k}) = \gamma \frac{x_{i,c,j,k} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}$$

$$\mu_i = \frac{1}{H} \frac{1}{CW} \sum_{c=1}^{C} \sum_{j=1}^{H} \sum_{k=1}^{W} x_{i,c,j,k}$$

→ selecting subset leads to group Normaliza.

summed out.

Basic

Multinomial distribution:

total n trial

Each trial : Possible k outcomes $\{E_1, E_2 \dots E_k\}$ with prob $\{P_1, \dots P_k\}$ respectively.

Lets assume $E_1 \xrightarrow{\text{happens}} n_1$, E_2 happens n_2 --- $E_k \rightarrow n_k$ times.

So, $n_1 + n_2 \dots + n_k = n$ // as n trial.

So, $$P = \frac{n!}{n_1!\, n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}$$

$$= \frac{n!}{n_1!\, n_2! \dots n_k!} \prod_{i=1}^{k} p_i^{n_i}$$

$$= n! \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!}$$

straight extentions.

Binomial distribution: Bernoulli trials.

n times fliping x positive $n-x$ negatives.

$$P(X = x \mid n, p) = {}^{n}C_{x}\, p^{x} (1-p)^{n-x}$$

$\downarrow$ (Distribution over Distribution)

Beta distribution (Prior to Dirichlet Distribution)

what if the p is a distribution itself ??

$p \in [0, 1]$

$\boxed{\text{Beta is conjugate prior for Binomial}}$

Now, $$P(p \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha - 1} (1-p)^{\beta - 1}$$

$$B(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)} = \binom{\alpha - 1}{\alpha + \beta - 2}$$ // continuous estimation.

Conjugate prior:

for some likelihood function, if we choose certain prior the posterior ends up being the same function → then conjugate prior.

Dirichlet Distribution:

Extended from multinomial probs. Conjugate prior to multinomial

what about the probs of multinomials $P_1 \dots P_k$??

① $\sum P_i = 1$ // already known, as condition each events prob needs to be 1.

$$P(P = \{P_i\} \mid \alpha_i) = \frac{\prod_i \Gamma(\alpha_i)}{\prod \Gamma(\sum_i \alpha_i)} \prod_i P_i^{\alpha_i - 1}$$

↳ to multinomial.

→ Generalization of Beta.

→ Distribution over multinomials.

→ Conjugate prior of multinomial

↳ given the data the $\{P_i\}$ will also be Dirichlet distribution.

* Beta/Gamma functions are different:

$$\text{Beta}(x, y) = B(x, y) = \frac{\Gamma(x)\,\Gamma(y)}{\Gamma(x+y)}$$ // binomial coeffs.

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx \ ; \ \Re(z) > 0$$

$\Gamma(z) = (z-1)!$ if z is a natural number ≥ 0 positive integer.