September 2nd

Assignment-01
CS772

MLE for λ:

given $p(x|\lambda) = \dfrac{\lambda^x \exp(-\lambda)}{x!}$

for $x_i \Rightarrow \log(p(x_i|\lambda)) = x_i \log(\lambda) - \lambda - \log(x_i!)$

for $x = (x_1 \cdots x_N) \Rightarrow \log(p(x|\lambda))$

$$NLL(\lambda) = \log(\lambda)\sum_{i=1}^{N} x_i - N\lambda - \sum_{i=1}^{N} \log(x_i!)$$

for MLE,

$$\frac{\partial NLL(\lambda)}{\partial \lambda} = 0$$

$$\Rightarrow \frac{\sum x_i}{\lambda} = N \Rightarrow \boxed{\lambda = \frac{\sum_{i=1}^{N} x_i}{N}}$$

MLE estimate

for MAP estimate,

$$L(\lambda) = MLE + \log(p(\lambda))$$
$$= \left(\sum x_i\right)\log\lambda - N\lambda - \sum \log(x_i!)$$
$$+ (\alpha-1)\log(\lambda) - \beta\lambda + \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)$$

$$\frac{\partial L(\lambda)}{\partial \lambda} = 0 \Rightarrow \frac{\sum x_i}{\lambda} - N + \frac{\alpha-1}{\lambda} - \beta = 0$$

$$\Rightarrow \boxed{\lambda_{MAP} = \frac{\sum x_i + (\alpha - 1)}{N + \beta}}$$

a kind of sum of obs. in prior

symbolically # obs. in prior

Since, γ-distribution is conjugate prior to Poisson, the posterior is γ-dist. as shown:

Posterior ∝ Prior · Likelihood

$$P(\lambda|x) \propto P(\lambda)\, P(x|\lambda)$$

$$P(\lambda) = \frac{\beta^{\alpha}\lambda^{\alpha-1}\exp(-\beta\lambda)}{\Gamma(\alpha)} \qquad P(x|\lambda) = \frac{\lambda^{\sum x_i}\exp(-N\lambda)}{\Pi(x_i)!}$$

from x_i's being i.i.d

$$\therefore\ P(\lambda|x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)\cdot K\,\Pi(x_i)!}\,\lambda^{\sum x_i+\alpha-1}\exp(-N\lambda-\beta\lambda)$$

K → Normalizing constant

$$\text{where } K = \int_0^{\infty}\frac{\beta^{\alpha}}{\Gamma(\alpha)\Pi(x_i)!}\lambda^{\sum x_i+\alpha-1}\exp(-N\lambda-\beta\lambda)\,d\lambda$$

$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)\Pi(x_i)!}\int_0^{\infty}\lambda^{\sum x_i+\alpha-1}\exp(-N\lambda-\beta\lambda)\,d\lambda$$

$$(*) \quad \therefore\ P(\lambda|x) = \frac{\lambda^{\sum x_i+\alpha-1}\exp(-N\lambda-\beta\lambda)}{\int_0^{\infty}\lambda^{\sum x_i+\alpha-1}\exp(-N\lambda-\beta\lambda)\,d\lambda}$$

Observe that $\text{Gamma}(\lambda; \sum x_i+\alpha, N+\beta)$ is the form of numerator i.e.

$$\int_0^{\infty}\text{Gamma}(\lambda; \sum x_i+\alpha, N+\beta) = 1$$

$$\Rightarrow \int_0^{\infty}\lambda^{\sum x_i+\alpha-1}\exp(-N\lambda-\beta\lambda) = \frac{\Gamma(\sum x_i+\alpha)}{(N+\beta)^{\sum x_i+\alpha}}$$

Putting Back in (*) we see,

Full posterior distribution → $\boxed{P(\lambda|x) \sim \text{Gamma}(\lambda; \sum x_i+\alpha, N+\beta)}$

(C) Mode of posterior $= \dfrac{\alpha + \sum_{i=1}^{N} x_i - 1}{N+\beta}$ (from given properties)

Priors mode $= \dfrac{\alpha - 1}{\beta}$ and $\lambda_{MLE} = \dfrac{\sum x_i}{N}$

So, clearly $\dfrac{N}{N+\beta}\lambda_{MLE} = \dfrac{\sum x_i}{N+\beta}$

$+ \quad \dfrac{\beta}{N+\beta} \text{Mode}_{prior} = \dfrac{\alpha - 1}{N+\beta}$

$$\underbrace{\frac{N}{N+\beta}\lambda_{MLE} + \frac{\beta}{N+\beta}\text{mode}_{prior}}_{\text{Linear combination}} = \text{Mode}_{posterior}.$$

$\mu_{prior} = \dfrac{\alpha}{\beta}$, $\mu_{posterior} = \dfrac{\alpha + \sum x_i}{N+\beta}$, $\lambda_{MLE} = \dfrac{\sum x_i}{N}$

$$\therefore \underbrace{\frac{N}{N+\beta}\lambda_{MLE} + \frac{\beta}{\beta+N}\mu_{prior}}_{\text{Linear combination}} = \mu_{posterior}$$

P2. From given formula,

$$\Sigma_{n+1} = \left(\beta \sum_{n=1}^{N+1} x_n x_n^T + \lambda I\right)^{-1} = \left(\underbrace{\beta \sum_{n=1}^{N} x_n x_n^T + \lambda I}_{\Sigma_N^{-1}} + x_{n+1} x_{n+1}^T\right)^{-1}$$

$$= \left(\Sigma_N^{-1} + x_{n+1} x_{n+1}^T\right)^{-1}$$

Putting it in given identity,

$$\left(\Sigma^{-1} + x_{n+1}x_{n+1}^T\right)^{-1} = \Sigma' - \frac{(\Sigma' x_n)(x_n^T \Sigma')}{(1 + x_n^T \Sigma' x_n)}$$

Now, $\sigma_{n+1}^2(x^*) = \beta^{-1} + x_*^T \Sigma'_{n+1} x_*$

$$= \beta^{-1} + x_*^T \left(\Sigma_n - \frac{(\Sigma'_n x_n)(x_n^T \Sigma'_n)}{1 + x_n^T \Sigma'_n x_n}\right) x_*$$

$$\sigma_{n+1}^2(x^*) = \underbrace{\beta^{-1} + x_*^T \Sigma'_n x_*}_{\sigma_n^2(x^*)} - x_*^T \frac{(\Sigma'_n x_n)(x_n^T \Sigma'_n)}{1 + x_n^T \Sigma'_n x_n} x_*$$

Relation for impact of data:

$$\boxed{\sigma_{n+1}^2(x^*) = \sigma_n^2(x^*) - x_*^T \underbrace{\frac{(\Sigma'_n x_n)(x_n^T \Sigma'_n)}{1 + x_n^T \Sigma'_n x_n} x_*}_{\Delta_n}}$$

Now, Σ'_n is a p.s.d. matrix (covar matrix)

$\therefore \forall v \in V, \; u^T \Sigma'_n u \geq 0$

or, $x_n^T \Sigma'_n x_n \geq 0 \Rightarrow$ denominator ≥ 1

Now if $(x_*^T \Sigma'_n x_n)$ is q then

$x_n^T \Sigma'^T_n x_* \text{ is } q^T = x_n^T \Sigma'_n x_*$ ← symmetric matrix

$$\therefore \Delta_n = \frac{q q^T}{\underbrace{1 + x_n^T \Sigma'_n x_n}_{\geq 1}}$$

Also, q is scalar, (from dimension)

$\therefore q q^T = q^2$

$$\therefore \Delta_n = \frac{q^2}{1 + x_n^T \Sigma'_n x_n} \geq 0 .$$

so, since $\Delta n \geq 0$

$$\therefore \sigma_{n+1}^2(x^*) = \sigma_n^2(x^*) - \Delta n$$
$$\leq \sigma_n^2(x^*)$$

Thus, variance decreases for a new observation with data (as expected).

P3

Class-distributions:

For real x_{nd}: $P(x_n | y_n)$: Gaussian
params: Σ_n, μ_n.

For binary x_{nd}: $P(x_n | y_n = 1/0)$: Bernoulli
params: bias (p)

for V-valued x_{nd}: $P(x_n / y_n = v')$: Multinoulli
Params: $P_{d,k}$: (prob. of x_n being as d given, y_n is k)

Formally,

for real x_{nd}, Total params: 2DK (D values for each of K classes)
- 2DK → DK means and DK variances.
+(K−1) → prior prob. over $(y_n = K)$

$$\therefore P(x_{nd} | y_n = k) = N(\mu_{d,k}, \sigma^2_{d,k})$$

Define an indicator Variable,

$$z_K(y_n) = \begin{cases} 1 & \text{if } y_n = k \\ 0 & \text{o/w} \end{cases}$$

$\therefore$

$$N_K = \sum_{n=1}^{N} z_K(y_n)$$

examples of class K.

$\therefore$ MLE estimators are,

Params for x_{nd}

$$\left\{ \begin{aligned} \mu_{d,K} &= \frac{\sum_{n=1}^{N} x_{nK}\, z_K(y_n)}{N_K} \\ \sigma^2_{d,K} &= \frac{\sum_{n=1}^{N} (x_{n,d} - \mu_{d,K})^2 z_K(y_n)}{N_K}. \end{aligned} \right\}$$

and, for Priors

$$\boxed{P(y_n = k \mid \pi) = \pi_K = \frac{N_K}{\sum N_K}}$$

= fraction of examples in class K.

(*) If x_{nd} is binary,

$$P(x_{nd} = 1 \mid y_n = K) = P_{d,K}$$

$\therefore$ there are a total of DK parameters.

and also $(k-1)$ params for priors of eq $(y_n = K)$

Aagain, $z_K(y_n) = \begin{cases} 1 & \text{if } y_n = K \\ 0 & \text{o/w} \end{cases}$

$$N_K = \sum_{n=1}^{k} z_K(y_n)$$

So, $\boxed{\pi_K = \frac{z_K(y_n)}{\sum z_K(y_n)}}$ ~~and MLE for $P_{d,K}$~~ $\boxed{P_{d,K} = \frac{\sum_{n=1}^{N} x_{nd}\, z_K(y_n)}{N_K}}$

So, as in (i) for case#1 the estimates are,

$$\boxed{P_{d,k} = \frac{\sum_{n=1}^{N} x_{nd}\, z_k(y_n)}{N_k} \quad , \quad \pi_k = \frac{N_k}{\sum_{k=1}^{K} N_k}}$$

(*) x_{nd} is V-valued,

$\therefore P(x_{nd} = i \mid y_{nd} = k) = P_{i,d,k}$ (Multinoulli dist).

with $\sum_{i=1}^{V} P_{i,d,k} = 1$.

$\therefore$ total params $= \underbrace{DK(V-1)}_{\text{for distribution on } n} + \underbrace{(k-1)}_{\text{for prior}}$

Again, $\pi_k = \frac{N_k}{\sum_{k=1}^{K} N_k}$, $P(y_n = k \mid \pi) = \pi_k$

Define,

$$z_{v,d}(x_n) = \begin{cases} 1 & \text{if } x_{nd} = v \\ 0 & \text{o/w} \end{cases}$$

MLE estimate (based on other answers)

$$\boxed{P_{i,d,k} = \frac{\sum_{n=1}^{N} z_{v,d}(x_n)\, z_k(y_n)}{N_k}}$$

P4

$K=3$ is the best model because it has the max. marginal likelihood

$P(y|K=3) = 2.5e{-10} > P(y|K = \text{any other value})$

$\therefore$ $K=3$ is good since,

$$\frac{P(y|M_1)}{P(y|M_2)} = \frac{P(M|y)}{M(M_2|y)} \frac{P(M_1)}{P(M_2)}$$

assuming $P(M_1) = P(M_2)$ higher $P(y|M)$ means better model.

If we can include a new point, it's better to include at $x=-4$, since the var. of uncertainty is max. at $x=-4$ for all $K=1,2,3$

at $K=3$, var at $x=-4$ is 7.180 (more than at all $x \in [-4,4]$)