Practice Set 2, Problem 1

$$y = \{y_1, y_2, \ldots, y_N\}$$

$$P(y|\lambda) = \frac{\lambda^{y_n} e^{-\lambda}}{y!} \quad \text{(Poisson)}$$

$$P(\lambda) = \frac{\beta^{\alpha}}{\Gamma\alpha} \lambda^{\alpha-1} e^{-\beta\lambda} \quad \text{(Gamma)}$$

① MLE and MAP :

$$\hat{\lambda}_{MLE} = \underset{\lambda}{\operatorname{argmax}} \sum_{n=1}^{N} \log P(y_n|\lambda) \quad (\underline{\text{i.i.d. data}})$$

$$\sum_{n=1}^{N} \log P(y_n|\lambda) = \left[\sum_{n=1}^{N} y_n \log\lambda\right] - N\cancel{\lambda +} \text{constant}$$

(terms that don't depend on λ)

Taking derivative and setting it to zero

$$\sum y_n \times \frac{1}{\lambda} - N = 0$$

$$\Rightarrow \boxed{\hat{\lambda}_{MLE} = \frac{\sum_{n=1}^{N} y_n}{N}}$$

MAP estimation will be almost identical with the extra $\log P(\lambda)$ term.

$$\hat{\lambda}_{MAP} = \underset{\lambda}{\operatorname{argmax}} \left[\sum_{n=1}^{N} \log P(y_n|\lambda) + \log P(\lambda)\right]$$

$$\log P(\lambda) = (\alpha-1)\log\lambda - \beta\lambda + \text{Constant}$$

terms that don't depend on λ

The MAP objective will be

$$\sum_{n=1}^{N} y_n \log\lambda - N\lambda + (\alpha-1)\log\lambda - \beta\lambda$$

Maximizing w.r.t. λ will give the MAP solution

$$\boxed{\hat{\lambda}_{MAP} = \frac{\sum_{n=1}^{N} y_n + \alpha - 1}{N+\beta}}$$

(2) Posterior distribution of λ

$$P(\lambda|y) = \frac{P(\lambda)P(y|\lambda)}{P(y)} = \frac{P(\lambda)\prod_{n=1}^{N} P(y_n|\lambda)}{P(y)}$$

Since the prior (gamma) and the likelihood (Poisson) ~~the~~ are conjugate, the posterior must be gamma!

Let's multiply the terms $P(\lambda)$ and $P(y_n|\lambda)$ are try to "identify" this gamma distribution's parameters:

$$P(\lambda|y) \propto P(y)\prod_{n=1}^{N} P(y_n|\lambda)$$

$$P(\lambda|y) \propto \lambda^{\alpha-1} e^{-\beta\lambda} \times \prod_{n=1}^{N} \frac{\lambda^{y_n} e^{-\lambda}}{y_n!}$$

$$\propto \lambda^{\sum_{n=1}^{N} y_n + \alpha - 1} e^{-(\beta+N)\lambda}$$

(ignoring the terms that don't depend on λ)

The above expression is clearly in form of a gamma distribution with

shape $= \sum_{n=1}^{N} y_n + \alpha$

rate $= \beta + N$

[No need to worry about constant of proportionality. It must be a gamma]

Thus $\boxed{P(y|\lambda) = \text{Gamma}\left(\sum_{n=1}^{N} y_n + \alpha, \beta + N\right)}$

(3) $$\hat{\lambda}_{MAP} = \frac{\sum_{n=1}^{N} y_n + \alpha - 1}{N+\beta} = \frac{\sum_{n=1}^{N} y_n}{N+\beta} + \frac{(\alpha-1)}{(N+\beta)}$$

$$= \underbrace{\left[\frac{\sum_{n=1}^{N} y_n}{N}\right]}_{\text{MLE}} \times \frac{N}{N+\beta} + \underbrace{\left[\frac{\alpha-1}{\beta}\right]}_{\text{Prior's mode}} \times \frac{\beta}{N+\beta}$$

Sum to 1 (Convex comb.)

Posterior's mean:

$$\frac{\sum_{n=1}^{N} y_n + \alpha}{N+\beta} = \underbrace{\frac{\sum_{n=1}^{N} y_n}{N}}_{\text{MLE}} \times \left[\frac{N}{N+\beta}\right] + \underbrace{\left[\frac{\alpha}{\beta}\right]}_{\text{Prior's mean}} \times \frac{\beta}{N+\beta}$$

④ $P(y_*|y) = \int P(y_*, \lambda|y)\, d\lambda$

$= \int P(y_*|\lambda) P(\lambda|y)\, d\lambda$

$\approx P(y_*|\hat{\lambda}_{MLE})$ (if using MLE)

or

$P(y_*|\hat{\lambda}_{MAP})$ (if using MAP)

In both these cases, $P(y_*|y)$ is simply a Poisson with parameters $\hat{\lambda}_{MLE}$ or $\hat{\lambda}_{MAP}$

If using the full posterior, which is $P(\lambda|y) = \text{Gamma}\left(\sum_{n=1}^{N} y_n + \alpha, \beta + N\right)$, $P(y_*|y)$ will be

$$P(y_*|y) = \int P(y_*|\lambda) P(\lambda|y)\, d\lambda$$

$$= \int \frac{\lambda^{y_*} e^{-\lambda}}{y_*!} \times \frac{(\beta+N)^{\sum_{n=1}^{N} y_n + \alpha}}{\Gamma\left(\sum_{n=1}^{N} y_n + \alpha\right)} \lambda^{\sum_{n=1}^{N} y_n + \alpha - 1} e^{-(\beta+N)\lambda}\, d\lambda$$

The above is actually a mixture of infinite many Poisson distributions. The result is actually not a Poisson but the "Negative Binomial" distribution.

(if you are interested in knowing more about this result, you may refer to the wikipedia article of NB distribution).

(This part was just for your "General knowledge" :-))

Scanned by CamScanner

Practice Set 2 (Problem 2)

$$\nabla_w NLL(w) = -\left[\sum_{n=1}^{N} y_n x_n - \frac{\exp(w^T x_n)}{1+\exp(w^T x_n)} \times x_n\right]$$

$P(y_n = 1 | w, x_n)$
$= \mu_n$

$$g = -\left[\sum_{n=1}^{N} y_n x_n - \mu_n x_n\right] = -\sum_{n=1}^{N} (y_n - \mu_n) x_n$$

The expression above can't be written by separating w on one side (like we did for linear regression). Therefore we can't find a closed form solution, and need iterative methods (e.g. gradient descent).

Intutive meaning of the gradient's expression :

$$g = -\sum_{n=1}^{N} (y_n - \mu_n) x_n$$

Contribution of x_n

$y_n = 0$ but $\mu_n \to 1$
or
$y_n = 1$ but $\mu_n \to 0$

if $y_n - \mu_n$ difference is large, which will happen if there is a LARGE MISPREDICTION, then x_n will contribute more to the gradient (we actually discussed this also while discussing about gradient descent).