Intro to Machine Learning (CS771A, Autumn 2018)
Practice Problem Set 2

Problem 1

Consider N count-valued observations $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$ drawn i.i.d. from a Poisson distribution $p(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$ where λ is the rate parameter of the Poisson. Assume a gamma prior on λ, i.e., $p(\lambda) = \text{Gamma}(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$, where $\alpha > 0$ is the *shape parameter* and $\beta > 0$ is the *rate parameter*, respectively, of the gamma.[1]. Note that, for this parameterization of gamma distribution, the prior's *mode* is $\frac{\alpha-1}{\beta}$ and mean is $\frac{\alpha}{\beta}$.

- Derive the MLE and MAP estimates for λ.
- Derive the posterior distribution for λ.
- Show that the MAP estimate (i.e., mode of the posterior) can be written as weighted combination of the MLE estimate and the prior's mode. Likewise, show that the posterior's *mean* can be written as a weighted combination of the MLE estimate and the prior's *mean*.
- Compute the predictive distribution $p(y_*|\boldsymbol{y})$, given the MLE and MAP estimate of λ, and also given the full posterior distribution over λ. In all the three cases, this would be an probability distribution over counts. Is it a Poisson in all the three cases?

Problem 2

The NLL for logistic regression model ($y_n \in \{0, 1\}$) was

$$\text{NLL}(\boldsymbol{w}) = -\sum_{n=1}^{N}(y_n \boldsymbol{w}^\top \boldsymbol{x}_n - \log(1 + \exp(\boldsymbol{w}^\top \boldsymbol{x}_n)))$$

Take its derivative w.r.t. $\boldsymbol{w}$ and show that the gradient can be written as a weighted combination of the N inputs. Also convince yourself that closed form solution for $\boldsymbol{w}$ is not possible in this case (unlike linear regression).

Take a close look at the form of the gradient expression you have obtained. This expression has an intuitive meaning in terms of which inputs contribute how much to each update of $\boldsymbol{w}$ when you apply the gradient descent procedure $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \boldsymbol{g}$, where $\boldsymbol{g}$ denotes the gradient.

[1] There is an alternate parameterization of gamma in terms of shape α and scale θ, for which $p(\lambda) \propto \lambda^{\alpha-1}e^{-\frac{\lambda}{\theta}}$