Convexity of Generalized Linear Models

Wednesday, August 12, 2020 2:21 PM

(a) [5 points] Derive an expression for the mean of the distribution. Show that $\mathbb{E}[Y \mid X; \theta]$ can be represented as the gradient of the log-partition function a with respect to the natural parameter η.

Hint: Start with observing that $\frac{\partial}{\partial \eta} \int p(y;\eta)dy = \int \frac{\partial}{\partial \eta} p(y;\eta)dy$.

Answer:

$$\int p(y;\eta)\,dy = 1$$

$$\frac{\partial}{\partial \eta} \int p(y;\eta)\,dy = 0$$

$$\int \frac{\partial}{\partial \eta} p(y;\eta)\,dy = 0$$

$$\int \frac{d}{\partial \eta} b(y) \exp(\eta y - a\eta)\,dy = 0$$

$$\int b(y) \exp(\eta y - a(\eta))(y - a'(\eta))\,dy = 0$$

$$\int (y\, p(y;\eta) - p(y;\eta)\, a'(\eta))\,dy = 0$$

$$\int y\, p(y;\eta) = \int p(y;\eta)\, a'(\eta)\,dy$$

$$E[y \mid x;\eta] = a'(\eta) \int p(y;\eta)\,dy$$

so, $E[Y \mid x;\eta] = a'(\eta)$

(b) [5 points] Next, derive an expression for the variance of the distribution. In particular, show that $\mathrm{Var}(Y \mid X; \theta)$ can be expressed as the derivative of the mean w.r.t η (*i.e.*, the second derivative of the log-partition function $a(\eta)$ w.r.t the natural parameter η.)

Answer:

$$\frac{\partial}{\partial^2 \eta} \int p(y;\eta)\,dy = \frac{\partial}{\partial \eta} \int (y - a'(\eta))\, b(y) \exp(\eta y - a(\eta))\,dy$$

$$= \int \frac{\partial}{\partial \eta} (\quad '' \quad)\,dy$$

$$= \int (y - a'(\eta))^2 b(y) \exp(\eta y - a\eta) - b(y) \exp(\eta y - a\eta)\, a''\eta$$

$$= \int b(y) \exp(\eta y - a(\eta)) \left[(y - a'(\eta))^2 - a''(\eta) \right] dy$$

$$= \int p(y;\eta) \left[(y - a'(\eta))^2 - a''(\eta) \right] dy$$

$$= \int p(y;\eta) (y - a'(\eta))^2 dy - \int p(y;\eta)\, a''(\eta)\, dy = 0$$

or/ $$\int p(y;\eta) [y^2 - 2y \underbrace{a'(\eta)}_{E[Y|x]} + a'(\eta)^2]\, dy = a''(\eta)$$

or/ $$\int p(y;\eta) y^2 dy - 2\, E[Y|x]^2 + E[Y|x]^2$$

$$E[Y^2 | x;\eta] - E[Y|x]^2 = a''(\eta)$$

$$\therefore \quad Var(Y | x; \theta) = a''(\eta)$$

(c) [5 points] Finally, write out the loss function $\ell(\theta)$, the NLL of the distribution, as a function of θ. Then, calculate the Hessian of the loss w.r.t θ, and show that it is always PSD. This concludes the proof that NLL loss of GLM is convex.

Hint: Use the chain rule of calculus along with the results of the previous parts to simplify your derivations.

Answer:

Remark: The main takeaways from this problem are:

- Any GLM model is convex in its model parameters.
- The exponential family of probability distributions are mathematically nice. Whereas calculating mean and variance of distributions in general involves integrals (hard), surprisingly we can calculate them using derivatives (easy) for exponential family.

$$\ell(\theta) = \sum_{i=1}^{m} \log (p(y^{(i)} | x^{(i)}; \theta)) = \sum_{i=1}^{m} \log p(y^{(i)};\eta)$$

$$\ell(\theta) = -\sum_{i=1}^{m} \log b(y^{(i)}) + y^{(i)} \theta^T x^{(i)} - a(\theta^T x^{(i)})$$

$$H_{jk} = \frac{\partial}{\partial\theta_j \partial\theta_k} \ell(\theta) = -\frac{\partial}{\partial\theta_j} \left(\sum_{i=1}^{m} y^{(i)} x_k^{(i)} - a'(\theta^T x^{(i)})\, x_k^{(i)} \right.$$

$$= \sum_{i=1}^{m} \left[a''(\theta^T x^{(i)})\, x_j^{(i)} x_k^{(i)} \right]$$

$\forall z,$

$$z^T H z = \sum_i \sum_j z_i H_{ij} z_j = \sum_i \sum_j z_i \, a''(\theta^T x)\, x_j x_k z_j$$

$$= a''(\theta^T x)\, (x^T z)^2$$

Since, $a''(\theta^T x)$ is $Var(y|x;\theta)$ so, $a''(\theta^T x)$ is non-negative

Hence $z^T H z \geq 0$ So ℓ is convex