Machine Learning and Computational Statistics
Generalized Linear Models and Gradient Boosting Machines

Intro: This document consists of concepts and exercises related to Generalized Linear Models and Gradient Boosting Machines. In the effort to model different distributions (e.g., Poisson Distribution, Exponential Distribution), there are two common approaches which are GLM and GBM. The goal is to make a good prediction of parameters (e.g., λ).

1. From the GLM approach, we predict $\lambda = \psi(w^T x)$ for some function $\psi$and some parameter vector $w \in \mathbf{R}^d$. Then write an expression for $\Pr_w(y|x)$, the predicted probability density function conditioned on x. Then, write the log-likelihood of $\Pr_w(D|x)$ which is L, and find the w_{MAP} by taking derivative of L.
2. From the GBM approach, believing that $w^T x$ does not extract enough information to predict λ, we set $\lambda = \psi(f(x)) = \exp^{f(x)}$. Then rewrite the objective function and apply the GBM process.
3. For conditional exponential distribution from GLM and GBM approaches, refer to conditional-exponential-distributions draft
4. For conditional poisson distribution from GLM and GBM approaches, refer to poisson gradient boosting
5. For Baysian linear regression from GLM approach, refer to Baysian linear regression

1 Conditional Exponential Distribution

The following note is an example of modeling exponential distribution from two approaches GLM and GBM.

Conditional exponential distributions from Generalized linear model (GLM) and Gradient Boosting Machine (GBM)

$\to$ Exp Dist $= \{ p_\lambda(y) = \lambda e^{-\lambda y} \mathbb{1}\{y \in [0,\infty)\} \mid \lambda \in (0,\infty) \}$

① GLM approach $\to$ predict $\lambda = \psi(w^T x)$ for some ψ, $w \in \mathbb{R}^d$.

Choose $\psi(\cdot) = \exp(\cdot)$ b/c $\exp(\cdot)$ is $\begin{cases} \text{monotonically } \uparrow \\ \text{differentiable} \end{cases}$

Thus:

$$p_w(y \mid x, w) = \lambda e^{-\lambda y}$$

$$= \begin{cases} \exp(w^T x)\, \exp(-\exp(w^T x) y) & \text{for } \lambda = \exp(w^T x),\ y \geq 0 \\ 0 & y < 0. \end{cases}$$

$$= \exp(w^T x) \exp(-e^{w^T x} y) \cdot \mathbb{1}\{y \geq 0\}$$

$D = ((x_1, y_1), \ldots, (x_n, y_n))$

$$L = p_w(D \mid x, w) = \prod_{i=1}^{n} p_w(y_i \mid x_i, w)$$

$$= \prod_{i=1}^{n} \exp(w^T x_i) \exp(-e^{w^T x_i} y_i) \quad \mathbb{1}\{y_i \geq 0\}$$

$L \neq 0 \iff y_i \geq 0 \quad \forall i \in \{1, \ldots, n\}$

$$\log L = \sum_{i=1}^{n} \left[\log(e^{w^T x_i}) + \log(e^{-e^{w^T x_i} y_i}) \right] = \sum_{i=1}^{n} w^T x_i - y_i e^{w^T x_i} = J(w)$$

$$w^* = \underset{w \in \mathbb{R}^d}{\operatorname{argmax}}\ J(w)$$

$$J(w) = \sum_{i=1}^{n} w^T x_i - y_i \exp(w^T x_i).$$

$J(w)$ is convex!

For any random (x_i, y_i).

$$J_i(w) = w^T x_i - y_i \exp(w^T x_i)$$

$$\Rightarrow \frac{\partial J_i(w)}{\partial w} = x_i - y_i \exp(w^T x_i)\, x_i .$$

$$\Rightarrow \text{SGD}: \quad w \leftarrow w + \eta \left[x_i - y_i \exp(w^T x_i)\, x_i \right]$$

(2) GBM approach : $\lambda = \psi(f(x))$ where f is some more general funct of x rather than setting $f(x) = w^T x$ as in GLM

$p(y_i | x_i f) = \exp(f(x)) \exp(-e^{f(x)} y_i)$

$\log p(y_i | x_i f) = f$

$$J(f) = \sum_{i=1}^{n} \ell(f(x_i), y_i).$$

$$p(y|\lambda) = \lambda \exp^{-\lambda y}.$$

$$p(y | x, f(x)) = \exp(f(x)) \exp(-\exp(f(x)) y) \qquad \lambda = \exp(f(x))$$

Key : $\ell(f(x_i), y_i) = \log(y_i | x_i, f(x_i))$

$$\log P(y | x, f(x)) = f(x) - y \exp(f(x)).$$

$$\Rightarrow \ell(f(x_i), y_i) = f(x_i) - y_i \exp(f(x_i)).$$

$$\Rightarrow r_i = \frac{\partial \ell(f(x_i), y_i)}{\partial f(x_i)} = 1 - y_i \exp(f(x_i)).$$

$$\Rightarrow h_m := \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{n} (-r_i - h(x_i))^2$$

$$= \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i \exp(f(x_i)) - 1 - h(x_i))^2$$

- The full GBM algorithm.
  - Set $f(x) = 0$.
  - for $m = 1$ to M:
    - Compute: $g_m = (1 - y_i \exp(f_{m-1}(x_i)))_{i=1}^{n}$
    - Fit regression model to $-g_m$.

$$h_m := \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{n} (-g_m^i - h(x_i))^2.$$

$$= \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i \exp(f_{m-1}(x_i)) - 1 - h(x_i))^2$$

-) Choose a fixed step size $\nu_m = \nu \in (0, 1]$

or

$$\nu_m = \operatorname*{argmax}_{\nu > 0} J(f_{m-1} + \nu h_m).$$

-) Take

$$f_m(x) = f_{m-1}(x) + \nu_m h_m(x).$$