Exercise 4

Sunday, 18 May 2025 10:32 PM

Exercise 4 : Logistic Regression

- **Task 1 :**

For a single observation (x_i, y_i), the Bernoulli model gives

$$p(y_i|x_i;w) = p_i^{y_i}(1-p_i)^{1-y_i}$$

where we will set $p_i = \sigma(w^T x_i)$

Joint likelihood for all N i.i.d obs.ⁿ,

$$L(w) = \prod_{i=1}^{N} p(y_i|x_i;w) = \prod_{i=1}^{n} (p_i)^{y_i}(1-p_i)^{1-y_i}$$

Taking log,

$$l(w) = \log L(w) = \sum_{i=1}^{N} [y_i \log p_i + (1-y_i)\log(1-p_i)]$$

As $p_i = \sigma(w^T x_i)$

$$l(w) = \sum_{i=1}^{N} [y_i \log[\sigma(w^T x_i)] + (1-y_i)\log(1-\sigma(w^T x_i))]$$

This is log-likelihood for log. regresⁿ on Bernoulli.

- **Task 2 :**

a] Convexity of negative log-likelihood

We know that log-likelihood is, its negtive would be,

$$-l(w) = \sum_{1}^{N} [-y_i \log[\sigma(w^T x_i)] - (1-y_i)\log(1-\sigma(w^T x_i))]$$

The grad. of each data pt. is,

$$\nabla_w [-l_i(w)] = [\sigma(w^T x_i) - y_i] x_i$$

Diff. again & we get hessian,

$$\nabla^2[-l(w)] = \sum_{1}^{N} \sigma(w^T x_i)[1-\sigma(w^T x_i)] x_i x_i^T$$

As $\sigma(w^T x_i)[1-\sigma(w^T x_i)] \geq 0$ & $x_i x_i^T$ is positive semi-definite, their sum is +ve semi-definite.

$\therefore \nabla^2[-l(w)] \geq 0 \Rightarrow l(w)$ is convex.

I believe convexity guarantees there are local minima. Any local min. is global. Thus, grad. descent will converge to global optimum (unique).

b] Binary cross-entropy loss:

We can see that,

$$\max_w \ell(w) \iff \min_w (-\ell(w)) \iff \min_w \frac{-1}{N} \ell(w)$$

Dividing by N turns negative log-likelihood into exactly the average binary cross-entropy:

$$\frac{-1}{N}\ell(w) = \frac{-1}{N}\sum_{i=1}^{N}\left[y_i \log \sigma(w^T x_i) + (1-y_i)\log(1-\sigma(w^T x_i))\right]$$
$$= L_{BCE}(w)$$

Thus,

$$\arg\max_w \ell(w) = \arg\min_w L_{BCE}(w)$$

showing that max.-likelihood estimation under Bernoulli GLM is equi. to minimizing the binary cross-entropy loss.

- Task 3:

a] Divergence of the MLE under linear separability.

The log-likelihood for $\{(x_i, y_i)\}_{i=1}^N$,

$$\ell(w) = \sum_{i}^{N}\left[y_i \log \sigma(w^T x_i) + (1-y_i)\log(1-\sigma(w^T x_i))\right]$$

where $\sigma(z) = 1/(1+e^{-z})$

By assumption there exists some weight w' s.t,

$$w'^T x_i \begin{cases} > 0 & \text{if } y_i = 1, \\ < 0 & \text{if } y_i = 0 \end{cases}$$

Let's take seq. $w_k = kw'$ — $[k = 1, 2, \ldots]$

Then for each i,

- if $y_i = 1$, $w_k^T x_i = k\,(w^T x_i) \longrightarrow +\infty$ so

  $\sigma(w_k^T x_i) \longrightarrow 1$ & $\log \sigma(w_k^T x_i) \longrightarrow 0$

- If $y = 0$, $w_k^T x_i = k(w_k^T x_i) \longrightarrow -\infty$ so

  $\sigma(w_k^T x_i) \longrightarrow 0$ & $\log(1 - \sigma(w_k^T x_i)) \longrightarrow 0$

∴ Every term in $l(w_k)$ approaches zero &

$$l(w_k) \longrightarrow 0 \quad \text{as} \quad k \longrightarrow \infty$$

For finite w, $l(w) < 0$. Thus supremum of log-likelihood is 0, and it attained only in the limit $\|w\| \longrightarrow \infty$

No finite max. exists ∴ MLE fails to converge under linear separability.

b] Mitigation:

In order to obtain optimization with a finite sol^n one typically penalizes large weights.

Introducing L_2 or L_1 penalty to neg. log likelihood.

$$\min_w \left[-l(w) + \lambda \|w\|_2^2\right] \text{ or } \min_w \left[-l(w) + \lambda \|w\|_1\right]$$

Also if we place a gaussian (for L_2) or Laplace (L_1) prior on w yeids a MAP estimate that remains finite even when the data are linearly separable. Regulari^n prevents weight mag. from diverging & often improves generalization.