Jake Schinto

Midterm

MLE and MAP

1.a. $P(w|D) = P(D|w) \cdot P(w)$ ← ignoring normalization

$P(D|w) = \prod_{i=1}^{m} \left(1/(1+e^{-(2y^{(i)}-1)\, w^T \cdot x^{(i)}}) \right)$ ← $(\{0,1\} \to \{-1,1\})$

$P(w) = \prod_{j=1}^{n} \frac{1}{2b} \cdot e^{-|w_j|/b}$

$P(w|D) = \prod_{i=1}^{m} \left(1/(1+e^{-(2y^{(i)}-1)\cdot w^T \cdot x^{(i)}}) \right) \cdot \prod_{j=1}^{n} \left(\frac{1}{2b} \cdot e^{\frac{-|w_j|}{b}} \right)$

equivalent ↓

$= \prod_{i=1}^{m} \left(y^{(i)} \left(1/(1+e^{-w^T \cdot x^{(i)}}) \right) + (1-y^{(i)}) \left(1 - \left(1/(1+e^{-w^T x^{(i)}}) \right) \right) \right) \cdot \prod_{j=1}^{n} \left(\frac{1}{2b} \cdot e^{\frac{-|w_j|}{b}} \right)$

1.b. $\ln(P(w|D)) = \ln(P(D|w)) + \ln(P(w))$

$\ln(P(D|w)) = \ln\left(P(y^{(i)}|x^{(i)}, w)\right)$

$= \sum_{i=1}^{m} \left(y^{(i)} \ln\left(P(y^{(i)}=1 | x^{(i)}, w)\right) + (1-y^{(i)}) \ln\left(1 - P(y^{(i)}=1 | x^{(i)}, w)\right) \right)$

$= \sum_{i=1}^{m} \left(y^{(i)} \left(w^T x^{(i)} \right) - \ln\left(1 + e^{w^T x^{(i)}}\right) \right)$

$\ln(P(w)) = \ln\left(\frac{1}{2b} \cdot e^{\frac{-|w_j|}{b}} \right) = \sum_{j=1}^{n} \ln(1) - \ln(2b) + \ln\left(e^{\frac{-|w_j|}{b}}\right)$

$= 0 - \ln(2b) + \left(\frac{-|w_j|}{b}\right)(1) = -\sum_{j=1}^{n} \left(\ln(2b) + \frac{|w_j|}{b} \right)$

when maximizing, $\ln(2b)$ is a constant and can be ignored because it won't change the max

$\rightarrow \ln(P(w)) \approx -\sum_{j=1}^{n} \frac{|w_j|}{b}$ ← $\lambda = b^{-1}$

So, $\ln(P(w|D)) \approx \sum_{i=1}^{m} \left(y^{(i)} \left(w^T x^{(i)} \right) - \ln\left(1 + e^{w^T x^{(i)}}\right) \right) - \sum_{j=1}^{n} \left(\ln(2b) + \frac{|w_j|}{b} \right)$

which is very similar to $-J(w)$:

where $L(w) = -\ln(P(D|w))$ and $R(w) \approx -\ln(P(w))$ + with $\lambda = b^{-1}$

therefore by maxing $\ln(P(w|D))$ it is the same as maxing $-J(w)$ = minimizing $J(w)$

Convex Analysis and Optimization

1. $f(w_j) = |w_j| \to |w_j|$ is a number ≥ 0

So, f'' is 0 which is nonnegative and therefore convex.

$$g(\alpha u + (1-\alpha)v) = \sum_{j=1}^{n} w_i f(\alpha u + (1-\alpha)v) \quad \leftarrow w_i = \lambda, \ \lambda = L^{-1} > 0$$

$$\leq \sum_{j=1}^{n} \lambda[\alpha f(u) + (1-\alpha) f(v)]$$

$$= \alpha \sum_{j=1}^{n} \lambda f(u) + (1-\alpha) \sum_{j=1}^{n} \lambda f(v)$$

$$= \alpha g(u) + (1-\alpha) g(v)$$

Therefore, $g(x) = \lambda \sum_{j=1}^{n} f(x)$ and is convex.
Since we know $L(x)$ is convex and (from HW3) we now know $\lambda R(x)$ is convex, then the sum of the two must also be convex.

2. when $x > 0$ $f(x) = x$ $f' = 1$
   when $x < 0$ $f(x) = -x$ $f' = -1$
   when $x = 0$ $f(x)$ is non differentiable

So, at $x=0$ it would be the set of all numbers between the two other functions connected to the point: $[-1, 1]$

← for all slopes $[-1,1]$ the linear underestimator can show it is convex

$$\partial F(x) = \begin{cases} \{1\} & x > 0 \\ \{-1\} & x < 0 \\ [-1,1] & x = 0 \end{cases}$$

3. $J(w) = L(w) + \lambda R(w)$

$\nabla J(w) = \nabla L(w) + \lambda \nabla R(w)$

$$L(w) = -\sum_{i=1}^{m}\left[y^{(i)}\left(w^T x^{(i)}\right) - \ln\left(1 + e^{w^T x^{(i)}}\right)\right]$$

$$w^T x^{(i)} = \sum_{j=1}^{n} w_j x^{(i)} \rightarrow \frac{d}{dw_j} = x^{(i)}$$

$$\nabla L(w) = -\sum_{i=1}^{m}\left[y^{(i)} x^{(i)} - \left(\frac{1}{1+e^{w^T x^{(i)}}}\right)\left(x^{(i)} e^{w^T x^{(i)}}\right)\right]$$

$$R(w) = \|w\|_1 = \sum_{j=1}^{n} |w_j| \rightarrow \frac{\partial}{\partial w_j} = \begin{cases} -1 & w_j < 0 \\ 1 & w_j > 0 \end{cases}$$

$$\nabla J(w) = -\sum_{i=1}^{m}\left[y^{(i)} x^{(i)} - \left(\frac{x^{(i)}}{1+e^{-w^T x^{(i)}}}\right)\right] \pm \lambda$$

$+$ if $w_j > 0$ $\quad$ $-$ if $w_j < 0$

4. $J(w) = L(w) + \lambda R(w)$

- From HW3, we proved that $L(w)$ is convex. From the definition of convex:
  $L(w + \mu e_j) \geq L(w) + \frac{dL}{dw_j}\mu$
- From Question 1, we proved that $R(w)$ is convex. By the definition of convex:
  $R(w + \mu e_j) \geq R(w) + \frac{dR}{dw_j}\mu$
- and multiply λ: $\lambda R(w + \mu e_j) \geq \lambda \cdot R(w) + \lambda \frac{dR}{dw_j}\mu$
- Combine the two inequalities:

$$L(w + \mu e_j) + \lambda R(w + \mu e_j) \geq L(w) + \lambda R(w) + \left(\frac{dL}{dw_j} + \frac{dR}{dw_j}\lambda\right)\mu$$

next page →

$$J(w+\mu e_j) \geq J(w) + \left(\frac{dL}{dw_j} + \lambda \frac{dR}{dw_j}\right)\mu$$

- let $g_j = \left(\frac{dL}{dw_j} + \lambda \frac{dR}{dw_j}\right) = \partial J(w)$
  - $J(w+\mu e_j) \geq J(w) + g_j \mu$ from def of sub gradient.
  - from Question 2 we see that $\frac{dR}{dw_j}$ at $w_j = 0$ is $[-1, 1]$.
  - Therefore, $\left|\frac{dR}{dw_j}\right| \leq 1$
  - let $a = \frac{dR}{dw_j}$
  - we get:

$$\partial J(w) = \left\{\frac{dL}{dw_j} + \lambda a : |a| \leq 1\right\}$$