1 Some properties of softmax classifier

$(x_i, y_i)_{i=1}^{N}$ $y_i \in [C]$ training set, $f: \mathcal{X} \to \mathbb{R}^C$

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} -\log\left(\frac{e^{f_{y_i}(x_i)}}{\sum_{j=1}^{c} e^{f_j(x_i)}}\right)$$

(a) If $\hat{R}_n(f) \leq \frac{\log 2}{n}$, then $\frac{1}{n}\sum_{i=1}^{n} 1(y_i \notin \arg\max_j f_j(x_i)) = 0$

<u>Proof.</u> Let $P_i = \dfrac{e^{f_{y_i}(x_i)}}{\sum_j e^{f_j(x_i)}}$ if x_i is classified correctly,

i.e. $y_i \in \arg\max_j f_j(x_i)$ $P_i = \dfrac{e^{f_{y_i}(x)}}{\sum_j e^{f_j(x_i)}} \geq \dfrac{e^{f_{y_i}(x)}}{C\, e^{f_{y_i}(x)}} = \dfrac{1}{C}$

if else, let $j_0 \in \arg\max_j f_j(x_i)$, we have $f_{y_i}(x) < f_{j_0}(x)$ so $P_i \leq \dfrac{e^{f_{y_i}(x)}}{e^{f_{y_i}(x)} + e^{f_{j_0}(x)}} < \dfrac{1}{2}$

Now $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} -\log P_i$. Let $I = \{i \in [C] \mid x_i$ is classified properly$\}$ $J = [C] - I$, $0 \leq |I| \leq C$

Then $\hat{R}_n(f) > \frac{1}{n}|J| \log 2$ (*). So $\frac{\log 2}{n} > \frac{1}{n}|J|\log 2$, $|J| < 1 \Rightarrow |J| = 0$

$I = [C]$ We are done.

(b) Show that $\frac{1}{n}\sum_{i=1}^{n} 1(y_i \notin \arg\max_j f_j(x_i)) < \frac{1}{\log 2}\hat{R}_n(f)$.

<u>Proof.</u> Following arguments in (a),

$$\frac{1}{n}\sum_{i=1}^{n} 1(y_i \notin \arg\max_j f_j(x_i)) = \frac{|J|}{n} < \frac{1}{\log 2}\hat{R}_n(f)$$

because of (*). We are done.

(c) If f classifies all data properly.

$$\lim_{\lambda\to\infty} \frac{\log(\hat{R}_n(\lambda f))}{\lambda} = -\min_{1\le i\le n} \min_{k\in\{j: f_j(x_i) < f_{y_i}(x_i)\}} [f_{y_i}(x_i) - f_k(x_i)]$$

Proof. According to the slides, f classifies all data implies $f_{y_i}(x_i) > \max_{j\neq y_i} f_j(x_i)$ thus when $\lambda\to\infty$ $\log(\hat{R}_n(\lambda))\to\infty$

We have $\log(\hat{R}_n(\lambda f)) = \log\left(\frac{1}{n}\sum_{i=1}^{n} -\log\frac{e^{\lambda f_{y_i}(x_i)}}{\sum_j e^{\lambda f_j(x_i)}}\right)$

According to L'Hopital law,

$$\lim_{\lambda\to\infty}\frac{\log(\hat{R}_n(\lambda f))}{\lambda} = \lim_{\lambda\to\infty}\frac{\hat{R}_n'(\lambda f)}{\hat{R}_n(\lambda f)} = \lim_{\lambda\to\infty}\frac{\hat{R}_n'(\lambda f)}{\frac{1}{n}\sum_{i=1}^{n} -\log\frac{e^{\lambda f_{y_i}(x_i)}}{\sum_j e^{\lambda f_j(x_i)}}}$$

$$= \lim_{\lambda\to\infty}\frac{\hat{R}_n'(\lambda f)}{\frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{e^{\lambda f_{y_i}(x_i)}}{\sum_j e^{\lambda f_j(x_i)}}\right)} = \lim_{\lambda\to\infty}\frac{-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{e^{\lambda f_{y_i}(x_i)}}{\sum_j e^{\lambda f_j(x_i)}}\right)'\frac{\sum_j e^{\lambda f_j(x_i)}}{e^{\lambda f_{y_i}(x_i)}}}{\text{Denominator}}$$

$$= \lim_{\lambda\to\infty}\frac{-\frac{1}{n}\sum_{i=1}^{n}\frac{f_{y_i}(x_i)e^{\lambda f_{y_i}(x_i)}\sum_j e^{\lambda f_j(x_i)} - e^{\lambda f_{y_i}(x_i)}\sum_j f_j(x_i)e^{\lambda f_j(x_i)}}{\left(\sum_j e^{\lambda f_j(x_i)}\right)e^{\lambda f_{y_i}(x_i)}}}{\text{Denominator}}$$

$$= \lim_{\lambda\to\infty}\frac{-\sum_{i=1}^{n}\frac{f_{y_i}(x_i)\sum_j e^{\lambda f_j(x_i)} - \sum_j f_j(x_i)e^{\lambda f_j(x_i)}}{\sum_j e^{\lambda f_j(x_i)}}}{\sum_{i=1}^{n}\frac{\sum_j e^{\lambda f_j(x_i)} - e^{\lambda f_{y_i}(x_i)}}{\sum_j e^{\lambda f_j(x_i)}}} = \lim_{\lambda\to\infty} -S(\lambda)$$

Because $\frac{\sum_j e^{\lambda f_j(x_i)} - e^{\lambda f_{y_i}(x_i)}}{\sum_j e^{\lambda f_j(x_i)}} \sim e^{-A_i}$, $\lambda\to\infty$, where $A_i = \min_{k\neq y_i}[f_{y_i}(x_i) - f_k(x_i)]$

and $\frac{f_y(x_i)\sum_j e^{\lambda f_j(x_i)} - \sum_j f_j(x_i)e^{\lambda f_j(x_i)}}{\sum_j e^{\lambda f_j(x_i)}} \sim A_i e^{-A_i}$, $\lambda\to\infty$

We have $-S(\lambda) \sim \frac{A e^{-A}}{e^{-A}}$, where $A = \min_i A_i$ (due to property of e^x)

i.e. $S(\lambda)\to -A$ when $\lambda\to\infty$. We are done.

2 Margin vs support vectors

$f(x;\theta) = \beta^T x + \beta_0$

$$\theta^* = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \ell(-y_i f(x_i;\theta)) + \frac{\lambda}{2}\|\beta\|_2^2$$

$\gamma^* = y_i f(x_i;\theta^*) = y_i(\beta^T x + \beta_0)$

(a) $\exists \alpha^* \in \mathbb{R}^n$, s.t. $\beta^* = \sum_{i=1}^{n} \alpha_i^* x_i$, and $|\alpha_i^*| \propto \ell'(-\gamma_i^*)$

Proof. The above problem is (Locally) unconstrained and differentiable.

Necessary condition is

$$\nabla_{\beta^*}\left(\frac{1}{n}\sum_{i=1}^{n} \ell(-y_i(\beta^T x + \beta_0)) + \frac{\lambda}{2}\|\beta\|_2^2\right) = 0$$

This is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n} \ell'(-\gamma_i^*)(-y_i x_i) + \lambda\beta = 0$$

i.e. $\beta = \sum_{i=1}^{n} \frac{\ell'(-\gamma_i^*) y_i}{n\lambda} x_i = \sum_{i=1}^{n} \alpha_i^* x_i$

$|\alpha_i^*| = \frac{\ell'(-\gamma_i^*)}{n\lambda} \propto \ell'(-\gamma_i^*)$ (as $y_i = \pm 1$) we are done

(b) when $\ell(t) = e^t$ $\ell'(t) = e^t$ according to (a), $|\alpha_i^*| \propto e^{-\gamma_i^*}$

This implies when γ_i^* is large (very confident to say it belongs to which class), the corresponding x_i contribute little (but larger than 0) to the formation of β^*

when $\ell(t) = \max_t(0, 1+t)$ when $\gamma_i^* > 1$, $\ell \equiv 0$ so $\ell' = 0 \Rightarrow \alpha_i^* = 0$

This implies confident points x_i (with margin $\gamma_i > 1$) doesn't contribute to formation of β^*

(c) The optimal β^* is built mainly with the information of (x_i, y_i), which are hard to classify (close to decision boundary), will never overfit the noise in confident points, so it generalizes well.

3 Derive a general soft-SVM

(a) Problem is $\min_{f\in\mathcal{F},\xi} \lambda\Omega(f)+\frac{1}{n}\sum_{i=1}^{n}t(\xi_i)$

$$\text{s.t. } y_if(x_i)\geqslant 1-\xi_i$$

$$\xi_i\geqslant 0$$

Because t is <u>penalization</u> function, WLOG let t be monotonically increasing.

Fix f, we see $\xi_i(f)=\begin{cases}1-y_if(x_i) & \text{if } y_if(x_i)<1\\ 0 & \text{if } y_if(x_i)\geqslant 1\end{cases}$ $=\max(0,1-y_if(x_i))$

Plugging it back, Problem becomes

$$\min_{f\in\mathcal{F}} \lambda\Omega(f)+\frac{1}{n}\sum_{i=1}^{n}t(\max(0,1-y_if(x_i))) \qquad (*)$$

We are done.

(b) Seen from $(*)$, choose $t(z)=z^2$ and we have

Problem is $\min_{f\in\mathcal{F}} \lambda\Omega(f)+\frac{1}{n}\sum_{i=1}^{n}\ell(y_if(x_i))$

where $\ell(z)=(\max(0,1-z))^2$

We are done.