Oscar Scholin
179 HW 2

1. $\sigma(x) = \dfrac{1}{1+e^{-x}}$

a) $\sigma'(x) = -(1+e^{-x})^{-2} \cdot -e^{-x} = \dfrac{e^{-x}}{(1+e^{-x})^2}$

$= \left(\dfrac{1}{1+e^{-x}}\right)\left(1 - \dfrac{1}{1+e^{-x}}\right)$

$= \sigma(x)(1-\sigma(x))$ $\square$

b) Gradient of log likelihood

Likelihood : $L(\theta) = P(x \mid \theta)$

log likelihood : $\ell(\theta) = \ln(L(\theta))$

assuming binary classification

We know $L(\theta) = \prod_i p_i^{y_i}(1-p_i)^{1-y_i}$, where y_i is the class label, p_i is the probability predicted by the model.

So $-\ell(\theta) = -\sum_i \left[y_i \ln(p_i) + (1-y_i)\ln(1-p_i)\right]$. Also, $p_i = \sigma(\underbrace{\vec{\theta}^T \vec{x}_i}_{\vec{\theta}\cdot\vec{x}_i})$

$\therefore \nabla_\theta(-\ell(\theta)) = -\sum_i y_i \dfrac{\sigma'(\vec{\theta}^T\vec{x}_i)}{\sigma(\vec{\theta}^T\vec{x}_i)} + (1-y_i)\dfrac{-\sigma'(\vec{\theta}^T\vec{x}_i)}{1-\sigma(\vec{\theta}^T\vec{x}_i)}$

since wrt θ

by part a

$= -\sum_i y_i(1-\sigma(\vec{\theta}^T\vec{x}_i))\vec{x}_i - (1-y_i)\sigma(\theta^T\vec{x}_i)\vec{x}_i$

$= -\sum_i y_i\vec{x}_i - y_i\sigma(\vec{\theta}^T\vec{x}_i)\vec{x}_i - \sigma(\theta^T\vec{x}_i)\vec{x}_i + y_i\sigma(\theta^T\vec{x}_i)\vec{x}_i$

$= \sum_i(\sigma(\vec{\theta}^T\vec{x}_i) - y_i)\vec{x}_i$

$= \sum_i(\xi_i - y_i)\vec{x}_i$ for $\xi_i = \sigma(\vec{\theta}^T\vec{x}_i)$

$= X^T(\vec{\xi} - \vec{y})$ (since the above is a dot product)

c) Compute the Hessian:

$H = \nabla_\theta(\nabla_\theta(-\ell))^T$

$= \nabla_\theta(X^T(\vec{\xi}-\vec{y}))^T$

independent of $\vec{\theta}$

$= \nabla_\theta(\vec{\xi}^T X - \vec{y}^T X)$

$= \nabla_\theta(\vec{\xi}^T X)$

using definition of $\vec{\xi}$ and generalizing the dot product

$= \nabla_\theta(\sigma(X\theta)^T X)$

$= X^T \mathrm{diag}(\vec{\xi}(1-\vec{\xi}))X = X^T S X$

$\therefore$ H is PSD $\iff$ $\text{diag}(\vec{\xi}(1-\vec{\xi}))$ PSD.

Note $\xi_i = \sigma(\vec{\theta}^T \vec{x}_i)$. $\therefore$ $\sigma(\theta^T \vec{x}_i)(1-\sigma(\vec{\theta}^T\vec{x})) \geq 0$ since range σ : $[0, 1]$.

$\therefore$ H is PSD.

2. Normalize Gaussian.

Compute $Z = \int_{-\infty}^{\infty} e^{-x^2/2\sigma^2} dx = \sqrt{\pi \frac{1}{\frac{1}{2\sigma^2}}} = \boxed{\sqrt{2\pi}\,\sigma}$. $\square$

using $\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}$

3. a) Start with

$$\underset{\vec{\omega}}{\arg\max} \sum_{i=1}^{N} \ln \mathcal{N}(y_i \mid \overbrace{\omega_0 + \vec{\omega}^T \vec{x}_i}^{\text{predicted } y_i}, \sigma^2) + \underbrace{\sum_{j=1}^{D} \ln \mathcal{N}(\omega_j \mid 0, \tau^2)}_{\text{log likelihood of these weights occuring}} .$$

Note the Gaussian prior is $$\mathbb{P}(\vec{\omega}) = \underbrace{\prod_j \mathcal{N}(\omega_j \mid 0, \tau^2)}_{\text{i.e. assume each weight comes from Gaussian centered @ 0 variance } \tau^2 \text{ by Occam's razor}}$$

In particular, note $\mathcal{N}(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Substituting in,

$$\underset{\vec{\omega}}{\arg\max} \sum_{i=1}^{N} \ln\left[\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(\omega_0 + \vec{\omega}^T \vec{x}_i)^2}{2\sigma^2}\right)\right] + \sum_{j=1}^{D} \ln\left[\frac{1}{\sqrt{2\pi}\,\tau} \exp\left(\frac{-\omega_j^2}{2\tau^2}\right)\right]$$

$$= \underset{\vec{\omega}}{\arg\max} \sum_{i=1}^{N} -\frac{(\omega_0 + \vec{\omega}^T \vec{x}_i)^2}{2\sigma^2} + \ln \frac{1}{\sqrt{2\pi}\,\sigma} + \sum_{j=1}^{D} \frac{-\omega_j^2}{2\tau^2} + \ln \frac{1}{\sqrt{2\pi}\,\tau}$$

$$= \underset{\vec{\omega}}{\arg\max} -\sum_{i=1}^{N} \frac{(\omega_0 + \vec{\omega}^T \vec{x}_i)^2}{2\sigma^2} + \ln \sqrt{2\pi}\,\sigma - \sum_{j=1}^{D} \frac{\omega_j^2}{2\tau^2} + \ln \sqrt{2\pi}\,\tau$$

$$= \underset{\vec{\omega}}{\arg\max} -\left[\underbrace{N \ln \sqrt{2\pi}\,\sigma + D \ln \sqrt{2\pi}\,\tau}_{\substack{\text{this term is constant} \\ \text{and so doesn't contribute} \\ \text{to the optimization}}} + \sum_{i=1}^{N} \frac{(\omega_0 + \vec{\omega}^T \vec{x}_i)^2}{2\sigma^2} + \sum_{j=1}^{D} \frac{\omega_j^2}{2\tau^2}\right]$$

Note $\underset{\vec{\omega}}{\arg\max} -(\ ) = \underset{\vec{\omega}}{\arg\min} (\ )$:

$$= \underset{\vec{\omega}}{\arg\min} \sum_{i=1}^{N} \frac{(\omega_0 + \vec{\omega}^T \vec{x}_i)^2}{2\sigma^2} + \sum_{j=1}^{D} \frac{\omega_j^2}{2\tau^2}$$

Rescale the problem by $2\sigma^2$ since this will also not affect optimization:

$$= \underset{\vec{\omega}}{\arg\min} \sum_{i=1}^{N} (\omega_0 + \vec{\omega}^T \vec{x}_i)^2 + \sum_{j=1}^{D} \underbrace{\frac{\sigma^2}{\tau^2}}_{\equiv \lambda} \omega_j^2 ,$$

which is Ridge Regression. □

b) Find solution $\vec{x}^*$ to Ridge Regression.

$$f \equiv \|A\vec{x} - \vec{b}\|^2 - \|\Gamma \vec{x}\|^2$$

Take gradient wrt $\vec{x}$ and set to 0:

$$\nabla_{\vec{x}} f = \nabla_{\vec{x}} \left[(A\vec{x} - \vec{b})^T (A\vec{x} - \vec{b}) + (\Gamma \vec{x})^T (\Gamma \vec{x}) \right]$$ writing out the norms

$$= \nabla_{\vec{x}} \left[(\vec{x}^T A^T - \vec{b}^T)(A\vec{x} - \vec{b}) + \vec{x}^T \Gamma^T \Gamma \vec{x} \right]$$

$$= \nabla_{\vec{x}} \left[\vec{x}^T A^T A \vec{x} \underbrace{- \vec{x}^T A^T \vec{b} - \vec{b}^T A \vec{x}}_{-2\vec{x}^T A^T \vec{b}} + \vec{b}^T \vec{b} + \vec{x}^T \Gamma^T \Gamma \vec{x} \right]$$

$$= 2A^T A \vec{x} - 2A^T \vec{b} + 2\Gamma^T \Gamma \vec{x}$$

Setting to 0:

$$2A^T A \vec{x}^* - 2A^T \vec{b} + 2\Gamma^T \Gamma \vec{x}^* = 0$$

$$(A^T A + \Gamma^T \Gamma)\vec{x}^* = A^T \vec{b}$$

$$\therefore \vec{x}^* = (A^T A + \Gamma^T \Gamma)^{-1} A^T \vec{b}$$

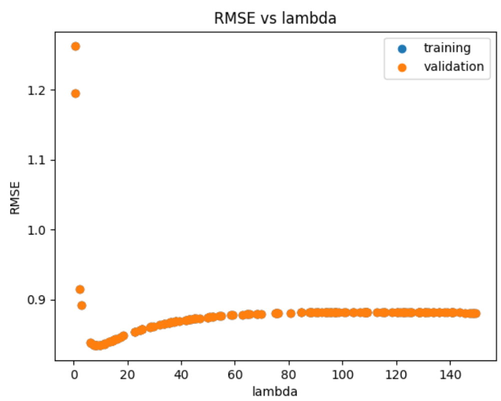
Let $\Gamma = \sqrt{\lambda} I \Rightarrow$ minimizes $f = ||A\vec{x} - \vec{b}||^2 + \lambda \vec{x}^T \vec{x}$

c) We have the following numerical results:

$\lambda^* = 8.6497$

RMSE on val: 0.8340

test: 0.8628



d) Solve: minimize $f = ||A\vec{x} + b\vec{1} - \vec{y}||^2 + ||\Gamma \vec{x}||^2$

$$\therefore f = (A\vec{x} + b\vec{1} - \vec{y})^T (A\vec{x} + b\vec{1} - \vec{y}) + (\Gamma\vec{x})^T (\Gamma \vec{x})$$

$$= (\vec{x}^T A^T + \vec{1}^T b - \vec{y}^T)(A\vec{x} + b\vec{1} - \vec{y}) + \vec{x}^T \Gamma^T \Gamma \vec{x}$$

$$= \vec{x}^T A^T A \vec{x} + 2b\vec{1}^T A \vec{x} - 2\vec{y}^T A \vec{x} - 2b\vec{1}^T \vec{y} + b^2 \overbrace{n}^{\vec{1}^T \vec{1}} + \vec{y}^T \vec{y} + \vec{x}^T \Gamma^T \Gamma \vec{x}$$

We must find gradient wrt $\vec{x}$ & b since these are fit params:

$$\nabla_{\vec{x}} f = 2A^T A \vec{x} + 2b A^T \vec{1} - 2\vec{y}^T A + 2\Gamma^T \Gamma \vec{x} \stackrel{!}{=} 0$$

$$\nabla_b f = 2\vec{1}^T A \vec{x} - 2\vec{1}^T \vec{y} + 2bn \stackrel{!}{=} 0$$

$$\therefore b^* = \frac{\vec{1}^T (\vec{y} - A\vec{x})}{n}$$

Now solve for $\vec{x}^*$:

$$A^T A \vec{x} + \Gamma^T \Gamma \vec{x} + \left(\frac{1^T(\bar{y} - A\vec{x})}{n}\right) A^T 1 - \bar{y} A^T \overset{!}{=} 0$$

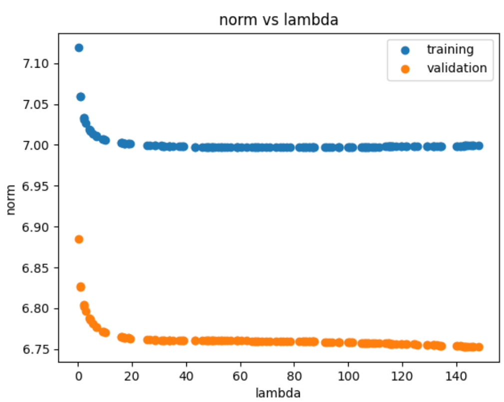
$$\left(A^T A + \Gamma^T \Gamma - \underbrace{\frac{1^T A}{n} A^T 1}_{\text{scalar, so can take transpose}}\right) \vec{x} + \frac{1^T \bar{y} A^T 1}{n} - \bar{y} A^T \overset{!}{=} 0$$

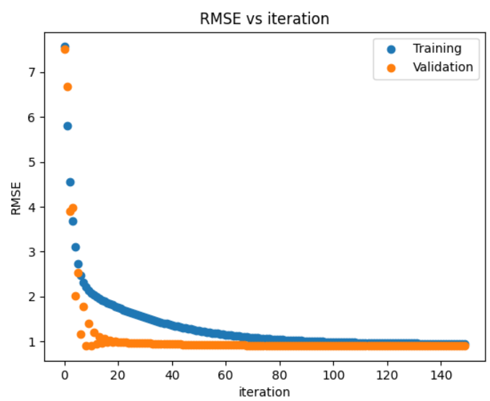
$$\left(A^T A + \Gamma^T \Gamma - \frac{1}{n} A^T 1^T 1 A\right) \vec{x} + \left(A^T \frac{1 1^T}{n} - A^T\right) \bar{y} = 0$$

$$A^T \left[\left(I - \frac{1}{n} 1 1^T\right) A + \Gamma^T \Gamma \right] \vec{x} = A^T \left(I - \frac{1}{n} 1 1^T\right) \bar{y}$$

$$\therefore \vec{x}^* = \left\{ A^T \left[\left(I - \frac{1}{n} 1 1^T\right) A + \Gamma^T \Gamma \right] \right\}^{-1} A^T \left(I - \frac{1}{n} 1 1^T\right) \bar{y}$$

e)





Gradient descent ↑

Diff in b: 2.909×10^{-11}

W: 7.996×10^{-1}

Diff in b: 1.539×10^{-1}

W: 7.996×10^{-1}