Deepak kumar

EECS 445 HW

1. (a) Batch gradient descent

$w = [2.4420, -2.8074]$ 531

(b) Stochastic gradient descent

$w = [2.4375, -2.8119]$ 519

(c) Newton's method

$w = [2.4464, -2.8164]$ 1

ii) In order of speed, the methods converged as follows:

Batch < Stochastic < Newton,

Batch - 531 iterations

Stochastic - 519 iterations

Newton - No iterations, 1 computation

(b) Plot given

ii) I would say a polynomial of degree 3-6 would best fit the data. There was evidence of overfitting for a polynomial of degree 10, when the E_{RMS} spiked up.

2(a) $$E_D(\vec{w}) = \frac{1}{2}\sum_{t=1}^{N} r^{(i)}(w^T x^{(i)} - t^{(i)})^2$$

Substitute $z = w^T x^{(i)} - t^{(i)}$

$$= E_D(\vec{w}) = \frac{1}{2}\sum_{i=1}^{N} r^{(i)} z^{(i)^2}$$

Transform $r^{(i)} \Rightarrow R$, diagonal matrix with $r^{(i)}$ on diagonal

$$E_D(\vec{w}) = \sum_{i=1}^{N} z^{(i)} R_{ii} z^{(i)} = z^T R z$$

$$= (w^T x - t)^T R (w^T x - t)$$

$$= (\Phi w - t)^T R (\Phi w - t), \text{ by matrix property rules.}$$

R is the diagonal matrix with values

$$R = \begin{cases} R_{ii} = \frac{1}{2} r^{(i)} \\ R_{ij} = 0 \end{cases}$$

(b) $(\Phi w - t)^T R (\Phi w - t)$

Before taking gradient, rewrite

$= (w^T \Phi^T - t^T)(R\Phi w - Rt)$

$E_D(\vec{w}) = w^T \Phi^T R \Phi w - t^T R \Phi w - w^T \Phi^T R t + t^T R t$

Now, take gradient wrt w^T

$\nabla_{w^T} E_D(\vec{w}) = \Phi^T R \Phi w - \Phi^T R t$

and set $= 0$, to maximize.

$\Phi^T R \Phi w - \Phi^T R t = 0$

$\Phi^T R \Phi w = \Phi R t$

$$\boxed{w = (\Phi^T R \Phi)^{-1} \Phi^T R t}$$

(c) $$P(t^{(i)} | x^{(i)}, w) = \frac{1}{\sqrt{2\pi}\,\sigma^{(i)}} \exp\left(\frac{-(t^{(i)} - w^T x^{(i)})^2}{2(\sigma^{(i)})^2}\right)$$

We have

$\log P(t^{(1)}, \ldots t^{(n)} | x, w)$

$= \log \prod_{i=1}^{N} \mathcal{N}(t^{(i)}; w^T \phi(x^{(i)}), ((\sigma^{(i)})^2)^{-1})$

$$= \sum_{i=1}^{N} \log\left(\sqrt{\frac{1}{\sqrt{2\pi}\sigma^{(i)2}}} \exp\left(-\frac{1}{2(\sigma^{(i)})^2}\{t^{(i)} - w^T(\phi(x^{(i)})\}^2\right)\right)$$

$$= \sum_{i=1}^{N} \left(\frac{1}{2}\log(\sigma^{(i)})^2 - \frac{1}{2}\log 2\pi - \frac{1}{(2\sigma^{(i)})^2}\{t^{(i)} - w^T(\phi(x^{(i)})\}^2\right)$$

2(c) Observe that the first two terms remain constants, so, simply put, we solve for $r^{(i)}$s in terms of $\sigma^{(i)}$s to be

$$r^{(i)} = \frac{1}{2(\sigma^{(i)})^2}.$$

(d) (iii)

When τ is too small or large, you have the problems of over/underfitting, where the curve it fits becomes too small/large or unreasonably tries to fit the data.

3 (c). kNN is a reasonable algorithm to use. Its advantages ~~is~~ are that it is simple and a reasonable approximation of a good classifier. It also uses local information which yields good behavior at optimal values of k.

The disadvantages involve large storage requirements w/ more features and data. Additionally, if k is poorly chosen, the classification accuracy decreases drastically, as adding too many neighbors ruins the probabilities.

4. (a) Start w/

$$\nabla_w \ell(w) = \sum_{n=1}^{N} (t^{(n)} - \sigma^{(n)})\, \phi(x^{(n)})$$

We know $\sigma^{(n)} = \sigma(w^T \phi(x^{(n)}))$, and that $\frac{\partial}{\partial s}\sigma(s) = \sigma(s)(1-\sigma(s))$

$= h(x^{(n)})$

Re-write gradient

$$\nabla_w \ell(w) = \sum_{n=1}^{N} (t^{(n)} - \sigma(w^T \phi(x^{(n)}))\, \phi(x^{(n)})$$

Take derivative wrt w

$$= \sum_{n=1}^{N} -\phi(x^{(n)})\, \sigma(w^T \phi(x^{(n)})(1-\sigma(w^T \phi(x^{(n)}))\, \phi(x^{(n)})$$

$$= -\sum_{n=1}^{N} -\phi(x^{(n)})\, \sigma^{(n)}(1-\sigma^{(n)})\, \phi(x^{(n)})$$

$$= -\sum_{n=1}^{N} \Phi_n h(x^{(n)})(1-h(x^{(n)}))\, \Phi_n^{(n)}$$

Thus, the hessian is

$$H = -\sum_{n=1}^{N} \Phi_n h(x^{(n)})(1-h(x^{(n)}))\, \Phi_n,$$ which is the desired form.

Now, we wts

$$z^T H z \leq 0$$

Start w/

$$\sum_i \sum_j z_i x_i x_j z_j = \sum_i z_i x_i \sum_j x_j z_j = (z^T x)(x^T z)$$

$$= (x^T z)^2 \geq 0 \quad \boxtimes$$

4(a) Now, look @ the form of the hessian.

$$H = -\sum_n \Phi_{ni}\, h(x^{(n)})\,(1-h(x^{(n)}))\,\Phi_{nj}$$

$$\therefore z^T H z = -z^T \sum_n \Phi_{ni}\, h(x^{(n)})\,(1-h(x^{(n)})\,\Phi_{nj}\, z$$

By our previous result, we thus see it is in the same form, so

$$-z^T \sum_n \Phi_{ni} h(x^{(n)})(1-h(x^{(n)}))\,\Phi_{nj} z \geq 0$$

or $z^T H z \leq 0$ ☒

(b) $w = [-1.4196, -0.7849, -0.1993]$

(c) Given plot.