HW # 4

1. $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \mu_2 = 5 \quad \Sigma_{11} = \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix} \quad \Sigma_{21}^T = \Sigma_{12} = \begin{bmatrix} 5 \\ 11 \end{bmatrix} \quad \Sigma_{22} = [14]$

(a) $p(x_1) = \mathcal{N}(\mu_1, \Sigma_{11}) = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix} \right)$

(b) $p(x_2) = \mathcal{N}(\mu_2, \Sigma_{22}) = \mathcal{N}(5, 14)$

(c) $p(x_1 | x_2) = \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$

where

$$\mu_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2) = \frac{1}{14} \begin{bmatrix} 5 \\ 11 \end{bmatrix} (x_2 - 5)$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} = \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix} - \frac{1}{14} \begin{bmatrix} 5 \\ 11 \end{bmatrix} [5 \quad 11]$$

$$= \begin{bmatrix} 59/14 & 57/14 \\ 57/14 & 61/14 \end{bmatrix}$$

(d) $p(x_2 | x_1) = \mathcal{N}(\mu_{2|1}, \Sigma_{2|1})$

$$\mu_{2|1} = \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (x_1 - \mu_1) = 5 + [5 \quad 11] \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix}^{-1} (x_1 - \mu_1)$$

$$= 5 + \begin{bmatrix} -\frac{23}{14} & \frac{13}{7} \end{bmatrix} x_1$$

$$\Sigma_{21} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} = 14 - [5 \quad 11] \begin{bmatrix} 6 & 8 \\ 8 & 13 \end{bmatrix}^{-1} \begin{bmatrix} 5 \\ 11 \end{bmatrix} = \frac{25}{14}$$

2.

(a) $P(y=1|x;\theta) = \sigma(\theta^T x)$ ← logistic model

Gaussian prior on weights, we have negative log likelihood

$$nll(\theta) = -\sum_i y_i \log \sigma(\theta^T x) + (1-y_i)\log(1-\sigma(\theta^T x)) + \frac{\lambda}{2}\|\theta\|_2^2.$$

Taking gradients,

$$\nabla_\theta l = \sum_i y_i (1-\sigma(\theta^T x))x - (1-y_i)\sigma(\theta^T x)x + \lambda\theta$$

$$= \sum_i [y_i - \sigma(\theta^T x_i)] x_i + \lambda\theta$$

$$= X^T(\sigma(X\theta) - y) + \lambda\theta$$

$$\nabla^2 l = \frac{d}{d\theta}\nabla l^T = \sum_i \nabla_\theta \sigma(\theta^T x) x_i^T + \lambda I$$

$$= X^T \operatorname{diag}[\sigma(X\theta)(1-\sigma(X\theta))] X + \lambda I$$

(Convergence plot attached)

We can see that Newton's Method is much faster than raw gradient descent.

(testing descriptions attached)

(b) For softmax regression we have:

$$\mathbb{P}(y = c | x, W) = \frac{1}{Z} \exp(W_c^T x)$$

$$= \frac{\exp(w_c^T x)}{\Sigma_i \exp(w_i^T x)}$$

Assuming Gaussian prior on each column of W:

$$nll(W) = -\log \prod_i \prod_c \mu_{ic}^{y_{ic}} - \lambda \, tr(W^T W)$$

$$= \sum_i \sum_c y_{ic} \log \mu_{ic} + \lambda \, tr(W^T W)$$

we can find:

$$\nabla_W \, nll = X^T(\mu - y) + \lambda W$$

$$y \in \{0,1\}^{n \times c}$$

$$y_{ij} \begin{cases} 1 \\ 0 \end{cases} \qquad y 1_c = 1_n$$

Similarly, we define $\mu \in [0,1]^{n \times c}$ as

$$\mu_i = S(x_i) = \frac{\exp(W^T x)}{1^T \exp(W^T x)}$$

Using stochastic gradient descent, we have test accuracies over diff. regularization param.

(accuracy vs λ attached)

(convergence plot attached)