

This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the `README.md` for this assignment includes instructions to regenerate this handout with your typeset L<sup>A</sup>T<sub>E</sub>X solutions.

---

1.a

Since  $g'(z) = g(z)(1 - g(z))$  and  $h(x) = g(\theta^T x)$ , it follows that  $\partial h(x)/\partial \theta_k = h(x)(1 - h(x))x_k$ .

Letting  $h_\theta(x^{(i)}) = g(\theta^T x^{(i)}) = 1/(1 + \exp(-\theta^T x^{(i)}))$ , we have

$$\frac{\partial \log h_\theta(x^{(i)})}{\partial \theta_k} =$$

$$\frac{\partial \log(1 - h_\theta(x^{(i)}))}{\partial \theta_k} =$$

Substituting into our equation for  $J(\theta)$ , we have

$$\frac{\partial J(\theta)}{\partial \theta_k} =$$

Consequently, the  $(k, l)$  entry of the Hessian is given by

$$H_{kl} = \frac{\partial^2 J(\theta)}{\partial \theta_k \partial \theta_l} =$$

Using the fact that  $X_{ij} = x_i x_j$  if and only if  $X = xx^T$ , we have

$$H =$$

To prove that  $H$  is positive semi-definite, show  $z^T H z \geq 0$  for all  $z \in \mathbb{R}^d$ .

$$z^T H z =$$

1.c

For shorthand, we let  $\mathcal{H} = \{\phi, \Sigma, \mu_0, \mu_1\}$  denote the parameters for the problem. Since the given formulae are conditioned on  $y$ , use Bayes rule to get:

$$\begin{aligned} p(y = 1|x; \mathcal{H}) &= \frac{p(x|y = 1; \mathcal{H})p(y = 1; \mathcal{H})}{p(x; \mathcal{H})} \\ &= \frac{p(x|y = 1; \mathcal{H})p(y = 1; \mathcal{H})}{p(x|y = 1; \mathcal{H})p(y = 1; \mathcal{H}) + p(x|y = 0; \mathcal{H})p(y = 0; \mathcal{H})} \\ &= \end{aligned}$$

1.d

First, derive the expression for the log-likelihood of the training data:

$$\begin{aligned}\ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^n p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi) \\ &= \sum_{i=1}^n \log p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) + \sum_{i=1}^n \log p(y^{(i)}; \phi) \\ &= \end{aligned}$$

Now, the likelihood is maximized by setting the derivative (or gradient) with respect to each of the parameters to zero.

**For  $\phi$ :**

$$\frac{\partial \ell}{\partial \phi} =$$

Setting this equal to zero and solving for  $\phi$  gives the maximum likelihood estimate.

**For  $\mu_0$ :**

**Hint:** Remember that  $\Sigma$  (and thus  $\Sigma^{-1}$ ) is symmetric.

$$\nabla_{\mu_0} \ell =$$

Setting this gradient to zero gives the maximum likelihood estimate for  $\mu_0$ .

**For  $\mu_1$ :**

**Hint:** Remember that  $\Sigma$  (and thus  $\Sigma^{-1}$ ) is symmetric.

$$\nabla_{\mu_1} \ell =$$

Setting this gradient to zero gives the maximum likelihood estimate for  $\mu_1$ .

For  $\Sigma$ , we find the gradient with respect to  $S = \Sigma^{-1}$  rather than  $\Sigma$  just to simplify the derivation (note that  $|S| = \frac{1}{|\Sigma|}$ ). You should convince yourself that the maximum likelihood estimate  $S_n$  found in this way would correspond to the actual maximum likelihood estimate  $\Sigma_n$  as  $S_n^{-1} = \Sigma_n$ .

**Hint:** You may need the following identities:

$$\begin{aligned}\nabla_S |S| &= |S| (S^{-1})^T \\ \nabla_S b_i^T S b_i &= \nabla_{Str} (b_i^T S b_i) = \nabla_{Str} (S b_i b_i^T) = b_i b_i^T \\ \nabla_S \ell &= \end{aligned}$$

Next, substitute  $\Sigma = S^{-1}$ . Setting this gradient to zero gives the required maximum likelihood estimate for  $\Sigma$ .

1.f

1.g

1.h

2.c

2.d



2.e