Problemset 1

$X \sim m \times n$, $y = 1 \times n$

1.

a) $\nabla^2_\theta J = \nabla_\theta \left[\frac{\partial J}{\partial g^i} \nabla_\theta g^i \right]$ ($\frac{\partial J}{\partial g^i} = x^i g(1-g)$, $g(\theta^T x^i) = \frac{1}{1+e^{-\theta^T x^i}}$) $= \nabla_\theta \left[\left(-\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{y^i}{g^i} + \frac{1-y^i}{1-g^i} \right) \times (1-g^i) g^i \right] \right]$

$= \nabla_\theta \left[+ \frac{X}{n} \sum_{i=1}^{n} \underbrace{-y^i (1-g^i) + (1-y^i) g^i}_{= g^i - y^i} \right] = \left(\frac{X X^T}{n} \sum_{i=1}^{n} g^i (1-g^i) \right) = H$ (∇_θ = $m \times 1$)

The Hessian is psd because $g(1-g) \in [0,1]$, $2y^i - 1 \geq 0$ and because XX^T is psd: $v^T X X^T v = \sum_{i=1}^{n} \langle v, x^i \rangle \langle x^i, v \rangle \geq 0$.

b) Done.

c) Baye's rule implies $p(y=1|x) = \frac{\overbrace{p(x|y=1) p(y=1)}^{= p(x, y=1)}}{\underbrace{p(x|y=0) p(y=0) + p(x|y=1) p(y=1)}_{= p(x)}}$. Plugging-in the above equations yields:

$= \frac{\phi \exp(-1/2 (x-\mu_1)^T \dots)}{\exp(-1/2 (x-\mu_0)^T \dots) \cdot (1-\phi) + \exp(-1/2 (x-\mu_1)^T \dots) \cdot \phi} = \frac{1}{\frac{1-\phi}{\phi} \cdot \exp\left[1/2 (x-\mu_1)^T \dots - 1/2 (x-\mu_0)^T \dots \right] + 1}$

$= \left[\exp\left(\log \frac{1-\phi}{\phi} - 1/2 \left(-x^T \Sigma^{-1} \mu_0 - \mu_0^T \Sigma^{-1} x + x^T \Sigma^{-1} \mu_1 + \mu_1^T \Sigma^{-1} x + \mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1 \right) \right) + 1 \right]^{-1}$

Define: $\theta_0 = \log \frac{\phi}{1-\phi} + \frac{1}{2} \left(-\mu_1^T \Sigma^{-1} \mu_1 + \mu_0^T \Sigma^{-1} \mu_0 \right)$ and $\theta^T = (-\mu_0 + \mu_1)^T \Sigma^{-1}$

$\Rightarrow p(y=1|x) = \left[\exp\left(-(\theta_0 + \theta^T x) \right) + 1 \right]^{-1}$, using that $x^T \mu_k = \mu_k^T x$.

The decision boundary (@threshold 0.5) has the eq $\theta^T x^1 + \theta_0 = 0$, which is clearly linear in x^1!

Homework Set 1 Cont

1d) $\phi, \mu_0, \mu_1, \Sigma = \arg\max \ell(\tilde{\phi}, \tilde{\mu}_0, \tilde{\mu}_1, \tilde{\Sigma})$ ∴ $p(x|y) = \frac{\exp\left[-\frac{1}{2}(x-\mu_y)^T \Sigma^{-1}(x-\mu_y)\right]}{(2\pi)^{d/2}|\Sigma|^{1/2}}$,

$\ell = \sum_{i=1}^{n} \log p(x|y;\dots) + \log p(y;\phi)$. $p(y) = \phi^y(1-\phi)^{1-y}$, $y \in \{0,1\}$

$$\ell \sim \sum_{i=1}^{n} -\frac{1}{2}\log|\Sigma| + \frac{1}{2}(x-\mu_{y^i})^T\Sigma^{-1}(x-\mu_{y^i}) + y^i\log\phi + (1-y^i)\log(1-\phi)$$

$$= -n\log|\Sigma| - \sum_{i=1}^{n}(x-\mu_{y^i})^T\Sigma^{-1}(x-\mu_{y^i}) + y^i\log\phi + (1-y^i)\log(1-\phi).$$

$$\frac{\partial \ell}{\partial \phi} = 0 \to \sum_{i=1}^{n} y^i/\phi = \sum_{i=1}^{n}\frac{1-y^i}{1-\phi} \iff (1-\phi)\sum_{i=1}^{n} y^i = (n - \sum_{i=1}^{n} y^i)\phi$$

$$\iff \sum_{i=1}^{n} y^i = n\phi \iff \underline{\phi = \frac{\sum_{i=1}^{n} y^i}{n} = \frac{\sum_{i=1}^{n} 1(y^i=1)}{n}}$$

$$\frac{\partial \ell}{\partial \mu_k} = 0 \to \sum_{i=1}^{n} 1\{y^i=k\}\cdot 2\Sigma^{-1}(x^i-\mu_k) = 0 \Rightarrow \underline{\mu_k = \frac{\sum_{i=1}^{n} 1\{y^i=k\}x^i}{\sum_{i=1}^{n} 1\{y^i=k\}}}$$

$\underline{k<k+(0,1)}$

For $\frac{\partial \ell}{\partial \Sigma}$, two important identities are necessary: $\log|\Sigma|^{-1} = (\Sigma^{-1})^T$ &

$(z^T\Sigma z) = \overbrace{-tr\,\Sigma^{-1}zz^T\Sigma^{-1}}^{=-\Sigma^{-1}zz^T\Sigma^{-1}}$. Using them, we get: $\frac{\partial \ell}{\partial \Sigma} = -n(\Sigma^{-1})^T + \sum_{i=1}^{n} tr\left(\Sigma^{-1}(x^i-\mu_{y^i})(x^i-\mu_{y^i})^T\Sigma^{-1}\right)$

Use that $\Sigma^T=\Sigma$ and that tr of a scalar = the scalar

$$\iff n\Sigma^{-1} = \sum_{i=1}^{n}\Sigma^{-1}(x^i-\mu_{y^i})(x^i-\mu_{y^i})^T\Sigma^{-1} \iff \underline{\Sigma = \sum_{i=1}^{n}(x^i-\mu_{y^i})(x^i-\mu_{y^i})^T/n}$$

Done..!

Homework set 1 cont

1f) Overall, the two decision boundaries look rather similar in some intercept and similar slope.

1g) Performance between GDA & log. reg. is practically identical for dataset 2. I'd say GDA performs ~~fairly~~ better on dataset 2, based on what I can see with my eye. This might be due to the data in dataset 2 actually coming from a two normal distributions, one for each class, thus satisfying the GDA assumptions. This notion is substantiated by the GDA & lin. regr. giving so similar results, lending credence to the sigmoid-description of the GDA hypothesis being correct. Further, to my eye, the data here does look Gaussian for both classes.

1h) Dataset 1 looks (with the eye) like $x_2 \sim \exp x_1$. Taking the log of x_2 and running GDA on the resulting data gives me a much better classifier, from what my eye can judge.

2a) Done. & 2b) Done.

c) $p(t^i=0|y^i=1,x^i) = p(y^i=1|t=0,x) \frac{p(t^i=0,x)}{p(y^i=1,x^i)} = \underline{0} \Rightarrow \underline{\underline{p(t^i=1|y^i=1,x^i)=1}}$

d) $p(t^i=1,x^i) = \underbrace{p(t^i=1|y^i=1,x^i)}_{=1} p(y^i=1) + \underbrace{p(t^i=1|y^i=0,x^i)}_{=p(y^i=0|t^i=1,x)} p(y^i=0,x^i) \frac{p(t^i=1,x^i)}{p(y^i=0,x^i)}$

$= p(y^i=1|x^i) + (1-\alpha) p(t^i=1|x^i)$

$\Leftrightarrow \underline{p(t^i=1|x^i) = \frac{1}{\alpha} p(y^i=1|x^i)}$

e) Ok... In other words, you are giving me $h(x^i)$ and assuming $p(t^i=1,x^i)$ = either 0 or 1, depending on x^i (ie wireless process), and you wish me to compute $\langle h(x^i), x^i | y^i=1 \rangle$ (probability of "correct positive").

Well, we know for a fact that there are no "false positives" here, so $h(x)=0$ if $t^i=0$. $h(x)|t^i=1 = \alpha$ from 2d.

$\Rightarrow E[h(x^i)|y^i=1] = E[p(y^i=1|x^i)|y^i=1] = E[\{\overbrace{p(y^i=1|t^i=1,x^i)}^{=\alpha} p(t^i=1,x^i) + \overbrace{p(y^i=1|t^i=0,x^i)}^{=0} p(t^i=0,x^i)\}|y^i=1] = E[\alpha p(t^i=1|x^i)|y^i=1]$

$= \alpha E[\overbrace{p(t^i=1|y^i=1,x^i)}^{=1}] = \underline{\underline{\alpha}}$

2f) Done.

3a) Exponential-family distributions have pdfs of the form $p(y;\eta) = b(y) e^{\eta^T T(y) - a(\eta)}$.
The Poisson-distribution is an example of these, with:
$b(y) = 1/y!$, $T(y) = y$, $\eta = \log\lambda$, $a(\eta) = \lambda = \exp(\eta)$.

b) $g(\eta) = E[y;\eta] = \langle p(y;\lambda) \rangle = \lambda = \exp(\eta)$

3c) $J(\theta) = \sum_{i=1}^{n} \log p(y^i | x^i; \theta) = \sum_{i=1}^{n} \log b(y^i) + \underbrace{\eta(x^i;\theta)}_{=\theta^T x^i} y^i - \underbrace{a(\eta(x^i;\theta))}_{=\exp(\eta)=\exp(\theta^T x^i)=h_\theta(x^i)}$

$\Rightarrow \frac{\partial l}{\partial \theta_j} = \sum_{i=1}^{n} x_j^i y^i - h_\theta(x^i) x_j^i$

learning rate, > 0.

$\Rightarrow \theta_j := \theta_j + \alpha \sum_{i=1}^{n} [y^i - h_\theta(x^i)] x_j^i \Leftrightarrow \theta := \theta + \alpha \sum_{i=1}^{n} [y^i - h_\theta(x^i)] x^i$

3d) Done.

4a) $\frac{\partial}{\partial \eta} \int p(y;\eta)\, dy = 0 = \int \frac{\partial p}{\partial \eta} dy = \int p(y;\eta) \cdot \frac{\partial [\eta y - a(\eta)]}{\partial \eta} dy$

$= \int p (y - \frac{\partial a}{\partial \eta}) dy \Leftrightarrow \underline{\langle y \rangle = E[Y;\eta] = \frac{\partial a}{\partial \eta}}$

4b) $\partial^2/\partial \eta^2 \int p\, dy = 0 = \int \frac{\partial}{\partial \eta} [p(y - \partial a/\partial \eta)]\, dy = \int \frac{\partial p}{\partial \eta}(y - \partial a/\partial \eta)$

$+ p(-\partial^2 a/\partial \eta^2)\, dy \Leftrightarrow \underline{\partial^2 a/\partial \eta^2} = \int p(y - \partial a/\partial \eta)^2 dy = \langle y^2 \rangle - \langle y \rangle^2 = \underline{Var[Y;\eta]}$

4c) $l(\theta) = \sum_{i=1}^{n} \log p(y^i | x^i; \theta) = \sum_{i=1}^{n} \log b(y^i) + \underbrace{\eta(x^i;\theta)}_{=\theta^T x^i} y^i - a[\eta(x^i;\theta)]$

$\Rightarrow \frac{\partial l}{\partial b_j} = \sum_{i=1}^{n} x_j^i y^i - \underbrace{\frac{\partial a}{\partial \eta}}_{=h_\theta(x^i)} x_j^i = \sum_{i=1}^{n} (y^i - \frac{\partial a}{\partial \eta}) x_j^i \Rightarrow \frac{\partial^2 l}{\partial \theta_k \partial \theta_j} = \sum_{i=1}^{n} \frac{\partial^2 a}{\partial \eta^2} x_k^i x_j^i \Leftrightarrow \nabla_\theta^2 l = \sum_{i=1}^{n} \frac{\partial^2 a}{\partial \eta^2} x^i x^{iT}$

Now, $u^T \nabla_\theta^2 l\, u = \sum_{i=1}^{n} \underbrace{\frac{\partial^2 a}{\partial \eta^2}(x^i;\theta)}_{\geq 0} (u^T x^i)(x^i u)$

$= \sum_{i=1}^{n} \underbrace{Var(Y|x^i;\theta)}_{\geq 0 \text{ always}} \cdot \underbrace{(\langle u, x \rangle)^2}_{\geq \text{always}} \geq 0 \ \forall u \Rightarrow \nabla_\theta^2 l$ is psd and thus $\underline{l(\theta)}$ is

a <u>convex function</u>, making it easy to minimise.

5a)

1) $\underbrace{\ell(\theta) = \sum_{i=1}^{n} \log p(y^i | x^i; \theta)}_{\text{log-loss}} = \sum_{i=1}^{n} + \log[(2\pi)^{1/2}\sigma] + \frac{1}{2\sigma^2}[y^{i2} - 2\overbrace{h_\theta(x^i)}^{=\theta^T x^i} y^i + h_\theta(x^i)^2]$

2) $\frac{\partial \ell}{\partial \theta_j} = \sum_{i=1}^{n} \frac{1}{2\sigma^2}(-2\hat{x}_j^i y^i + 2h_\theta(\hat{x}^i)\hat{x}_j^i) = -\frac{1}{\sigma^2}\sum_{i=1}^{n} \hat{x}_j^i (y^i - h_\theta(\hat{x}^i))$

$\Rightarrow \theta_j := \theta_j + \alpha \sum_{i=1}^{n} (y^i - h_\theta(\hat{x}^i))\hat{x}_j^i \Leftrightarrow \theta := \theta + \alpha \sum_{i=1}^{n} (y^i - \overbrace{h_\theta(\hat{x}^i)}^{=\theta^T \hat{x}^i})\hat{x}^i$

5b) Done.

5c) Using Taylor expansion, it is possible to ~~train~~ fit a sinusoid (which is what the training data is, along with some added noise) when the number of polynomials is sufficient.

With this in mind, we see ~~what~~ from the plot that k=1,2,3 is too few polynomials, 5&10 is sufficient, while 20 is too many (~~the~~ the model is ~~overfitting~~ fitting noise).

5d) Now (obviously) the k=1 model fits the best, ~~as one of its few~~ because it contains the correct feature $\sin(x)$, which matches the structure of the data perfectly, and little other features, making overfitting a non-issue. Predictably, adding more higher-order polynomials increases overfitting.

~~This fit is for better~~ This result is in strong contrast to 5c, where ~~mostly~~ a longish Taylor series is needed to correctly capture the fine-wave (albeit not too long, as otherwise overfitting occurs).

PS: It's very important that the added fine feature has the same wavenumber as the one ~~in the~~ which generated the data.

5e) We here see extreme overfitting: $k \geq 5$ fits exactly to the datapoints.