Name: Qiran Li Uni: ql2397

Problem1

a)

$$\hat{\pi} = \arg\max_{\pi} \sum_{i=1}^{n} \ln p(y_i|\pi) = \arg\max_{\pi} \sum_{i=1}^{n} \ln\left(\pi^{y_i}(1-\pi)^{(1-y_i)}\right)$$

$$= \arg\max_{\pi} \sum_{i=1}^{n} \left(y_i \ln\pi + (1-y_i)\ln(1-\pi)\right)$$

$$lik = \sum_{i=1}^{n} y_i \ln\pi + (1-y_i)\ln(1-\pi)$$

$$\frac{\partial lik}{\partial \pi} = \frac{1}{\pi}\sum_{i=1}^{n} y_i + \frac{1}{\pi-1}\sum_{i=1}^{n}(1-y_i) = \frac{(\pi-1)\sum y_i + \pi(n-\sum y_i)}{\pi(\pi-1)}$$

$$= \frac{n\pi - \sum y_i}{\pi(\pi-1)} = 0$$

So $\hat{\pi} = \dfrac{\sum_{i=1}^{n} y_i}{n}$

b) $Y = \{0,1\}$ $d = 1:D$

So

$$\hat{\lambda}_{y,d} = \arg\max_{\lambda_{y,d}} \sum_{d=1}^{D}\left(\ln p(\lambda_{y,d}) + \sum_{i=1}^{n} \ln p(x_{i,d}|\lambda_{y_i,d})\right)$$

$\lambda_{y,d} \overset{iid}{\sim} \text{Gamma}(2,1)$ $x_{i,d}|y_i \sim \text{Pois}(\lambda_{y_i,d})$ $y_i \overset{iid}{\sim} \text{Bern}(\pi)$

$$\ln p(\lambda_{y,d}) = \ln \lambda_{y,d} e^{-\lambda_{y,d}} = \ln\lambda_{y,d} - \lambda_{y,d}$$

$$\ln p(x_{i,d}|\lambda_{y_i,d}) = \ln\left(\frac{\lambda_{y_i,d}^{x_{i,d}} e^{-\lambda_{y_i,d}}}{x_{i,d}!}\right)$$

$$= x_{i,d}\ln\lambda_{y_i,d} - \lambda_{y_i,d} - \ln x_{i,d}!$$

$$lik = \sum_{d=1}^{D}\left(\ln\lambda_{y,d} - \lambda_{y,d} + \sum_{i=1}^{n} x_{i,d}\ln\lambda_{y_i,d} - \lambda_{y_i,d} - \ln(x_{i,d}!)\right)$$

Since for each $\lambda_{y_i,d}$, it can be either $\lambda_{0,d}$ or $\lambda_{1,d}$

Create an indicator variable $\mathbb{1}(y_i = y)$

$$lik = \sum_{d=1}^{D} \left(\ln \lambda_{y,d} - \lambda_{y,d} + \sum_{i=1}^{n} x_{i,d} \ln \lambda_{y,d} \mathbb{1}(y_i = y) - \lambda_{y_i,d} \mathbb{1}(y_i = y) - \ln(x_{i,d}!) \right)$$

$$\frac{\partial lik}{\partial \lambda_{y,d}} = \sum_{d=1}^{D} \left(\frac{1}{\lambda_{y,d}} - 1 + \frac{\mathbb{1}(y_i = y) \sum_{i=1}^{n} x_{i,d}}{\lambda_{y,d}} - n_y \right) = 0$$

So $\hat{\lambda}_{y,d} = \dfrac{1 + \mathbb{1}(y_i = y) \sum_{i=1}^{n} x_{i,d}}{n_y + 1}$ where $n_y = \sum_{i=1}^{n} \mathbb{1}(y_i = y)$

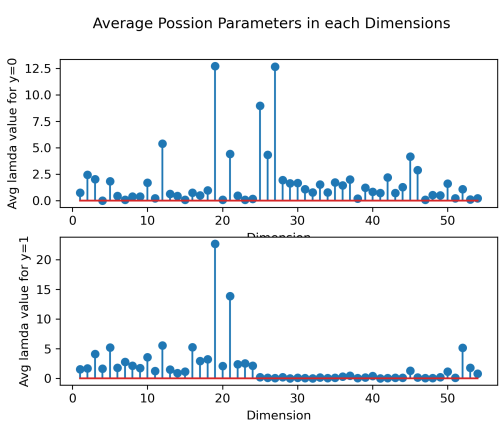
P2

a)

		y' predict by model	
		0	1
y ground truth	0	2300	487
	1	111	1702

So the accuracy is $\dfrac{2300 + 1702}{4600} = 87\%$

b)

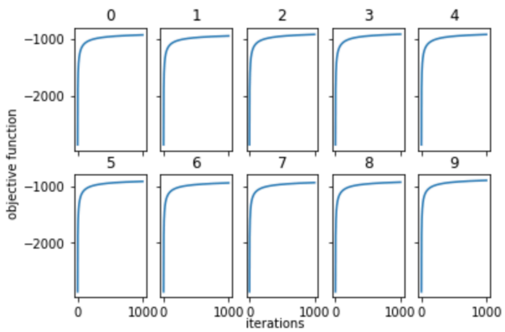


(16: free
52: !)

Dim 16 has a smaller λ value for nonspam email and a larger λ value for spam email.

Dim 52 has a larger λ value for spam email and a smaller λ value for nonspam email.

c)



d) $L'(w) = L(w) + (w-w_t)^T \nabla L(w_t) + \frac{1}{2}(w-w_t)^T \nabla^2 L(w_t)(w-w_t)$

set

$w_{t+1} = \arg\max_w L'(w)$

$L(w) = \ln P(y, w|X)$

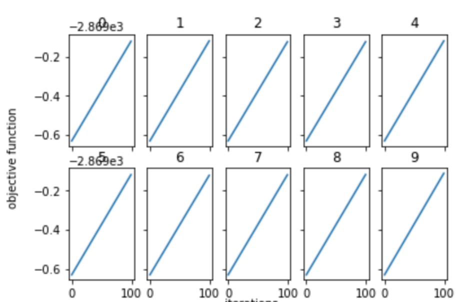
$\nabla^2 L(w_t) =$

$\nabla^2 \ln P(y, w_t|X) = -\lambda I - \sum_{i=1}^{n} \sigma_i(y_i \cdot w_t)(1-\sigma_i(y_i \cdot w_t)) x_i x_i^T$

So

$w_{t+1} = w_t - \eta (\nabla_w^2 L)^{-1} \nabla_w L$ where

$\nabla_w L = -\sum_{i=1}^{n} \sigma_i(w_t)(1-\sigma_i(w_t)) x_i x_i^T$



e)

		y' predict by model	
		-1	1
y ground truth	-1	2649	138
	1	290	1423

so the accuracy is $\frac{2649+1423}{4600} = 88.5\%$

P3

a)

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
5	1.96628	1.93314	1.92342	1.9222	1.92477	1.92921	1.93463	1.94058	1.94682	1.95321
7	1.92016	1.90488	1.90808	1.9159	1.9248	1.9337	1.94225	1.95038	1.95809	1.96544
9	1.89765	1.90252	1.91765	1.93251	1.9457	1.95723	1.9674	1.97649	1.98474	1.99234
11	1.89051	1.91498	1.93885	1.95794	1.97322	1.98576	1.99638	2.0056	2.01384	2.02134
13	1.89585	1.93559	1.9646	1.9855	2.00131	2.01388	2.02431	2.03331	2.04132	2.04864
15	1.9096	1.95955	1.9908	2.01192	2.02737	2.03947	2.04946	2.0581	2.06585	2.07298

b) $b=11$ $\sigma^2=0.1$ has the lowest $\widehat{RMSE}=1.89$.
It's better than in HW1 which the lowest RMSE is 2.2.
Drawback 1: ① Gaussian Process runs too slow. It's in general more computationally expensive
② Gaussian Process can learn training data really well, but ends up memorizing the data, which produces the problem of overfitting

c)