CS229 Problem Set 1

Tianyu Du

Monday 15^{th} July, 2019

1 Question 1: Linear Classifiers

1.1 Question 1(a)

Remark 1.1. Let $d \in \mathbb{N}$, then $[d]$ denotes the set $\{1, \dots, d\}$.

Lemma 1.1.

$$g'(z) = g(z)\,(1 - g(z)) \tag{1.1}$$

Lemma 1.2. For every $x, z \in \mathbb{R}^n$,

$$\sum_i \sum_j z_i x_i z_j x_j = (x^T z)^2 \geq 0 \tag{1.2}$$

$$\begin{aligned}
\nabla_\theta J(\theta) &= -\frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} \frac{g'(\theta^T x^{(i)})}{g(\theta^T x^{(i)})} - (1 - y^{(i)}) \frac{g'(\theta^T x^{(i)})}{1 - g(\theta^T x^{(i)})} \right) x^{(i)} && (1.3)\\
&= -\frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} (1 - g(\theta^T x^{(i)})) - (1 - y^{(i)}) g(\theta^T x^{(i)}) \right) x^{(i)} && (1.4)\\
&= -\frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - g(\theta^T x^{(i)}) \right) x^{(i)} && (1.5)\\
\implies \frac{\partial J(\theta)}{\partial \theta_j} &= -\frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - g(\theta^T x^{(i)}) \right) x_j^{(i)} \; \forall j \in [d] && (1.6)\\
\implies \forall k \in [d],\ \frac{\partial^2 J(\theta)}{\partial \theta_j \partial \theta_k} &= -\frac{1}{n} \sum_{i=1}^{n} \left(-g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)})) x_k^{(i)} \right) x_j^{(i)} && (1.7)\\
&= \frac{1}{n} \sum_{i=1}^{n} \left(g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)})) x_j^{(i)} x_k^{(i)} \right) && (1.8)
\end{aligned}$$

Therefore, $H_J(\theta)$ can be constructed from the array of second order derivatives of $J(\theta)$ as

$$H_J(\theta)_{j,k} := \frac{1}{n} \sum_{i=1}^{n} \left(g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)})) x_j^{(i)} x_k^{(i)} \right) \tag{1.9}$$

Notice that since $g(\theta^T x^{(i)}) \in (0, 1)$, therefore $g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)})) > 0$ for every θ and $x^{(i)}$.

Proof. Show that $H_J(\theta) \succeq 0$: let $z = (z_1, \dots, z_d) \in \mathbb{R}^d$, then

$$z^T H_J(\theta) \in \mathbb{R}^{1 \times d} \tag{1.10}$$

Then the β^{th} column of $z^T H_J(\theta)$ is

$$z^T H_J(\theta)_\beta = \frac{1}{n} \sum_{\alpha=1}^{d} \sum_{i=1}^{n} g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)})) z_\alpha x_\alpha^{(i)} x_\beta^{(i)} \tag{1.11}$$

Therefore

$$z^T H_J(\theta) z = \frac{1}{n} \sum_{\beta=1}^{d} \sum_{\alpha=1}^{d} \sum_{i=1}^{n} g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)})) z_\alpha x_\alpha^{(i)} x_\beta^{(i)} z_\beta \tag{1.12}$$

$$= \frac{1}{n} \sum_{i=1}^{n} g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)})) \sum_{\beta=1}^{d} \sum_{\alpha=1}^{d} z_\alpha x_\alpha^{(i)} x_\beta^{(i)} z_\beta \tag{1.13}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \underbrace{g(\theta^T x^{(i)})(1 - g(\theta^T x^{(i)}))}_{>0 \ \because g(\cdot)\in(0,1)} \underbrace{(z^T x)^2}_{\geq 0} \geq 0 \tag{1.14}$$

Hence, $H_J(\theta) \succeq 0$ is shown by showing $z^T H_J(\theta) z$ for every $z \in \mathbb{R}^d$. ■

1.2 Question 1(b) Coding

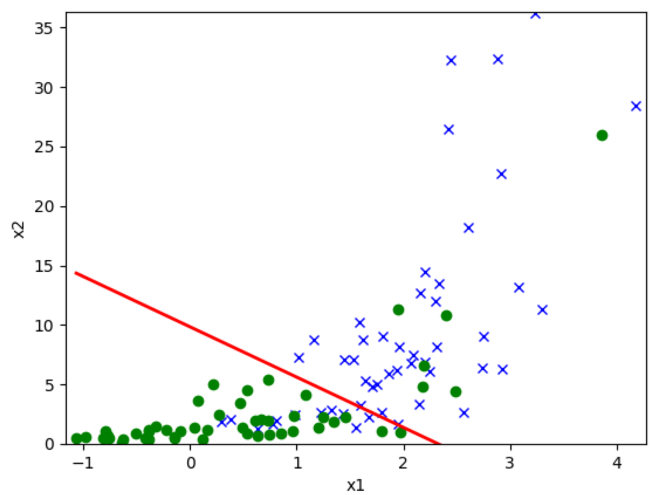

Figure 1: Logistic Regression Decision Boundary on Dataset 1

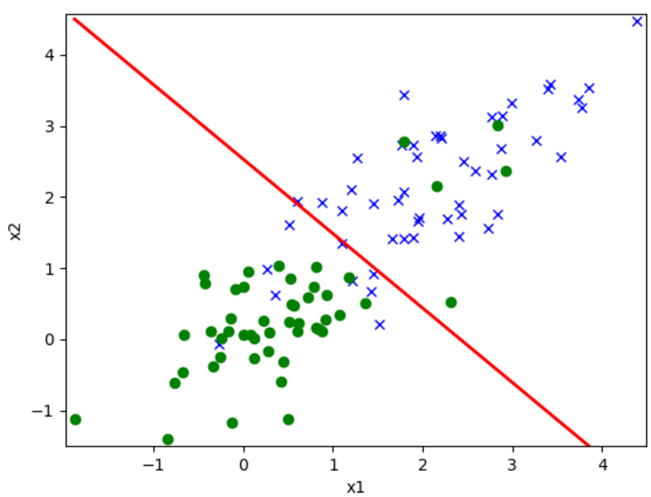

Figure 2: Logistic Regression Decision Boundary on Dataset 2

1.3 Question 1(c)

Proof. By Bayes' theorem,

$$p(y=1|x;\phi,\mu_0,\mu_1,\Sigma) = \frac{p(x|y=1;\phi,\mu_0,\mu_1,\Sigma)p(y=1;\phi,\mu_0,\mu_1,\Sigma)}{p(x;\phi,\mu_0,\mu_1,\Sigma)} \tag{1.15}$$

Define

$$z := \frac{p(x|y=1;\phi,\mu_0,\mu_1,\Sigma)p(y=1;\phi,\mu_0,\mu_1,\Sigma)}{p(x;\phi,\mu_0,\mu_1,\Sigma)} \tag{1.16}$$

$$\Theta := \{\phi,\mu_0,\mu_1,\Sigma\} \tag{1.17}$$

Conditioned on particular x, y is either 0 or 1, therefore,

$$p(y=0|x;\Theta) = 1-z \tag{1.18}$$

$$\implies \frac{z}{1-z} = \frac{p(y=1|x;\Theta)}{p(y=0|x;\Theta)} \tag{1.19}$$

$$= \frac{p(x|y=1;\Theta)p(y=1;\Theta)}{p(x|y=0;\Theta)p(y=0;\Theta)} \tag{1.20}$$

$$= \frac{\phi}{1-\phi}\frac{\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)}{\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)} \tag{1.21}$$

$$\implies \log\frac{z}{1-z} = \log\frac{\phi}{1-\phi} \tag{1.22}$$

$$+\left(-\frac{1}{2}x^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}x - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right) \tag{1.23}$$

$$-\left(-\frac{1}{2}x^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}x - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0\right) \tag{1.24}$$

$$= \log\frac{\phi}{1-\phi} + \left((\mu_1-\mu_0)^T\Sigma^{-1}x + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right) \tag{1.25}$$

$$\implies \frac{z}{1-z} = \exp\underbrace{\left(\log\frac{\phi}{1-\phi} + \left((\mu_1-\mu_0)^T\Sigma^{-1}x + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right)\right)}_{=:\Delta} \tag{1.26}$$

$$\implies z = \frac{\exp(\Delta)}{1+\exp(\Delta)} = \frac{1}{1+\exp(-\Delta)} \tag{1.27}$$

Therefore

$$\frac{p(x|y=1;\phi,\mu_0,\mu_1,\Sigma)p(y=1;\phi,\mu_0,\mu_1,\Sigma)}{p(x;\phi,\mu_0,\mu_1,\Sigma)} = \frac{1}{1+\exp(-(\theta^Tx+\theta_0))} \tag{1.28}$$

where

$$\theta = (\Sigma^{-1})^T(\mu_1-\mu_0) \tag{1.29}$$

$$\theta_0 = \log\frac{\phi}{1-\phi} + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 \tag{1.30}$$

1.4 Question 1(d)

1.4.1 ϕ

Proof.

$$\frac{\partial}{\partial \phi}\ell(\phi, \cdot) = \frac{\partial}{\partial \phi}\sum_{i=1}^{n} \underbrace{\log p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma)}_{\perp \phi} + \log p(y^{(i)}; \phi) \tag{1.31}$$

$$= \frac{\partial}{\partial \phi}\sum_{i=1}^{n} \log p(y^{(i)}; \phi) \tag{1.32}$$

$$= \frac{\partial}{\partial \phi}\sum_{i=1}^{n} \log \phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}} \tag{1.33}$$

$$= \frac{\partial}{\partial \phi}\sum_{i=1}^{n} y^{(i)} \log \phi + (1-y^{(i)}) \log(1-\phi) \tag{1.34}$$

$$= \sum_{i=1}^{n} y^{(i)}\frac{1}{\phi} - (1-y^{(i)})\frac{1}{1-\phi} \tag{1.35}$$

The first order condition of maximizing likelihood becomes

$$\sum_{i=1}^{n} y^{(i)}\frac{1}{\phi} - (1-y^{(i)})\frac{1}{1-\phi} = 0 \tag{1.36}$$

$$\implies \sum_{i=1} \frac{y^{(i)}}{\phi} + \frac{y^{(i)}}{1-\phi} - \frac{1}{1-\phi} = 0 \tag{1.37}$$

$$\implies \sum_{i=1}^{n} y^{(i)}\frac{1-\phi+\phi}{\phi(1-\phi)} - \frac{1}{1-\phi} = 0 \tag{1.38}$$

$$\implies \sum_{i=1}^{n} y^{(i)}\frac{1}{\phi(1-\phi)} = n\frac{1}{1-\phi} \tag{1.39}$$

$$\implies \phi = \frac{1}{n}\sum_{i=1}^{n} y^{(i)} \tag{1.40}$$

■

1.4.2 μ_0

Proof.

$$\frac{\partial}{\partial \mu_0}\ell(\mu_0,\cdot) = \frac{\partial}{\partial \mu_0}\sum_{i=1}^{n}\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) + \underbrace{\log p(y^{(i)};\phi)}_{\perp \mu_0} \tag{1.41}$$

$$= \frac{\partial}{\partial \mu_0}\sum_{i=1}^{n}\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) \tag{1.42}$$

$$= \frac{\partial}{\partial \mu_0}\sum_{i=1}^{n}\Bigg\{\overbrace{y^{(i)}\log\left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right)\right]}^{\perp \mu_0} \tag{1.43}$$

$$+ (1-y^{(i)})\log\left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right)\right]\Bigg\} \tag{1.44}$$

$$= \frac{\partial}{\partial \mu_0}\sum_{i=1}^{n}(1-y^{(i)})\left(\underbrace{\log\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}}_{\perp \mu_0} - \frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right) \tag{1.45}$$

$$= \frac{\partial}{\partial \mu_0}(-1)\sum_{i=1}^{n}(1-y^{(i)})\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0) = 0 \tag{1.46}$$

$$\implies \sum_{i=1}^{n}(1-y^{(i)})\Sigma^{-1}(x^{(i)}-\mu_0) = 0 \tag{1.47}$$

$$\tag{1.48}$$

$$\implies \sum_{i=1}^{n}\Sigma^{-1}(1-y^{(i)})x^{(i)} = \sum_{i=1}^{n}\Sigma^{-1}(1-y^{(i)})\mu_0 \tag{1.49}$$

$$\implies \sum_{i=1}^{n}(1-y^{(i)})x^{(i)} = \sum_{i=1}^{n}(1-y^{(i)})\mu_0 \tag{1.50}$$

$$\implies \mu_0 = \frac{\sum_{i=1}^{n}(1-y^{(i)})x^{(i)}}{\sum_{i=1}^{n}(1-y^{(i)})} = \frac{\sum_{i=1}^{n}\mathbb{1}\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^{n}\mathbb{1}\{y^{(i)}=0\}} \tag{1.51}$$

■

1.4.3 μ_1

Proof.

$$\frac{\partial}{\partial \mu_1}\ell(\mu_1,\cdot) = \frac{\partial}{\partial \mu_1}\sum_{i=1}^{n} \log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) + \underbrace{\log p(y^{(i)};\phi)}_{\perp \mu_1} \tag{1.52}$$

$$= \frac{\partial}{\partial \mu_1}\sum_{i=1}^{n} \log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) \tag{1.53}$$

$$= \frac{\partial}{\partial \mu_1}\sum_{i=1}^{n} \Big\{ y^{(i)} \log\left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right)\right] \tag{1.54}$$

$$+ \underbrace{(1-y^{(i)}) \log\left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right)\right] \Big\}}_{\perp \mu_1} \tag{1.55}$$

$$= \frac{\partial}{\partial \mu_1}\sum_{i=1}^{n} y^{(i)} \left(\underbrace{\log \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}}_{\perp \mu_1} - \frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right) \tag{1.56}$$

$$= \frac{\partial}{\partial \mu_1}(-1)\sum_{i=1}^{n} y^{(i)}\frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1) = 0 \tag{1.57}$$

$$\implies \sum_{i=1}^{n} y^{(i)}\Sigma^{-1}(x^{(i)}-\mu_1) = 0 \tag{1.58}$$

$$\implies \sum_{i=1}^{n} \Sigma^{-1}y^{(i)}x^{(i)} = \sum_{i=1}^{n} \Sigma^{-1}y^{(i)}\mu_1 \tag{1.59}$$

$$\implies \sum_{i=1}^{n} y^{(i)}x^{(i)} = \sum_{i=1}^{n} y^{(i)}\mu_1 \tag{1.60}$$

$$\implies \mu_1 = \frac{\sum_{i=1}^{n} y^{(i)}x^{(i)}}{\sum_{i=1}^{n} y^{(i)}} = \frac{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)}=1\}} \tag{1.61}$$

■

1.4.4 Σ^{-1}

Lemma 1.3. Given matrices A, B such that AB and BA are squared, then

$$\operatorname{tr}(AB) = \operatorname{tr}(BA) \tag{1.62}$$

Proof. In lecture notes. ■

Lemma 1.4.

$$\frac{\partial}{\partial A}\operatorname{tr}(AB) = \frac{\partial}{\partial A}\operatorname{tr}(BA) = B^T \tag{1.63}$$

Proof. Let $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times n}$. Let $a_i \in \mathbb{R}^m$ denote the i^{th} row of matrix A, let $b_j \in \mathbb{R}^m$ denote the j^{th} column of matrix B,

$$\begin{aligned}
\operatorname{tr}(AB) &= \operatorname{tr}\begin{pmatrix} a_1\cdot b_1 & a_1\cdot b_2 & \cdots & a_1\cdot b_n \\ a_2\cdot b_1 & a_2\cdot b_2 & \cdots & a_2\cdot b_n \\ \vdots & & \ddots & \vdots \\ a_n\cdot b_1 & a_n\cdot b_2 & \cdots & a_n\cdot b_n \end{pmatrix} && (1.64)\\
&= \sum_{i=1}^{n} a_i\cdot b_i && (1.65)\\
&= \sum_{i=1}^{n} A_{i,j}B_{j,i} && (1.66)\\
&\implies \frac{\partial \operatorname{tr}(AB)}{\partial A_{i,j}} = B_{j,i} && (1.67)\\
&\implies \nabla_A \operatorname{tr}(AB) = B^T && (1.68)
\end{aligned}$$

Since $\operatorname{tr}(AB) = \operatorname{tr}(BA)$, $\nabla_A \operatorname{tr}(BA) = B^T$ as well. ■

Proof.

$$\begin{aligned}
&\frac{\partial}{\partial\Sigma}\ell(\Sigma,\cdot) = 0 && (1.69)\\
&\implies \frac{\partial}{\partial\Sigma}\sum_{i=1}^{n}\log(p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)) + \underbrace{\log(p(y^{(i)};\phi))}_{\perp\Sigma} = 0 && (1.70)\\
&\iff \frac{\partial}{\partial\Sigma}\sum_{i=1}^{n}\log(p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)) = 0 && (1.71)\\
&\implies \frac{\partial}{\partial\Sigma}\sum_{i=1}^{n} y^{(i)}\left[-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)\right] && (1.72)\\
&+ (1-y^{(i)})\left[-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)\right] = 0 && (1.73)
\end{aligned}$$

where $-\frac{d}{2}\log(2\pi)$ is constant and independent from Σ, therefore it can be dropped in the first order condition.

$$\implies \frac{\partial}{\partial \Sigma} \sum_{i=1}^{n} y^{(i)} \left[-\frac{1}{2} \log(|\Sigma|) - \frac{1}{2} (x^{(i)} - \mu_1)^T \Sigma^{-1} (x^{(i)} - \mu_1) \right] \tag{1.74}$$

$$+ (1 - y^{(i)}) \left[-\frac{1}{2} \log(|\Sigma|) - \frac{1}{2} (x^{(i)} - \mu_0)^T \Sigma^{-1} (x^{(i)} - \mu_0) \right] = 0 \tag{1.75}$$

$$\implies \frac{\partial}{\partial \Sigma} \sum_{i=1}^{n} -\frac{1}{2} \log(|\Sigma|) - \frac{y^{(i)}}{2} (x^{(i)} - \mu_1)^T \Sigma^{-1} (x^{(i)} - \mu_1) - \frac{1 - y^{(i)}}{2} (x^{(i)} - \mu_0)^T \Sigma^{-1} (x^{(i)} - \mu_0) = 0 \tag{1.76}$$

$$\implies \frac{\partial}{\partial \Sigma} \sum_{i=1}^{n} \log(|\Sigma|) + y^{(i)} (x^{(i)} - \mu_1)^T \Sigma^{-1} (x^{(i)} - \mu_1) + (1 - y^{(i)}) (x^{(i)} - \mu_0)^T \Sigma^{-1} (x^{(i)} - \mu_0) = 0 \tag{1.77}$$

$$\implies n\Sigma^{-1} + \frac{\partial}{\partial \Sigma} \sum_{i=1}^{n} y^{(i)} \mathrm{tr}((x^{(i)} - \mu_1)^T \Sigma^{-1} (x^{(i)} - \mu_1)) + (1 - y^{(i)}) \mathrm{tr}((x^{(i)} - \mu_0)^T \Sigma^{-1} (x^{(i)} - \mu_0)) = 0 \tag{1.78}$$

$$\implies n\Sigma^{-1} + \frac{\partial}{\partial \Sigma} \sum_{i=1}^{n} y^{(i)} \mathrm{tr}(\Sigma^{-1} (x^{(i)} - \mu_1)(x^{(i)} - \mu_1)^T) + (1 - y^{(i)}) \mathrm{tr}(\Sigma^{-1} (x^{(i)} - \mu_0)(x^{(i)} - \mu_0)^T) = 0 \tag{1.79}$$

$$\implies n\Sigma^{-1} + \sum_{i=1}^{n} y^{(i)} (x^{(i)} - \mu_1)(x^{(i)} - \mu_1)^T + (1 - y^{(i)}) (x^{(i)} - \mu_0)(x^{(i)} - \mu_0)^T = 0 \tag{1.80}$$

$$\implies n\Sigma^{-1} + \sum_{i=1}^{n} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T = 0 \tag{1.81}$$

$$\implies \Sigma = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \in \mathbb{R}^{d \times d} \tag{1.82}$$

$\blacksquare$

1.5 Question 1(e) Coding

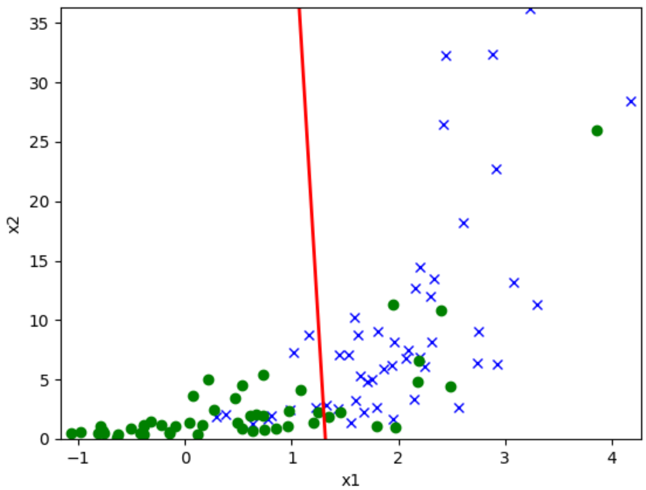

Figure 3: GDA Decision Boundary on Dataset 1

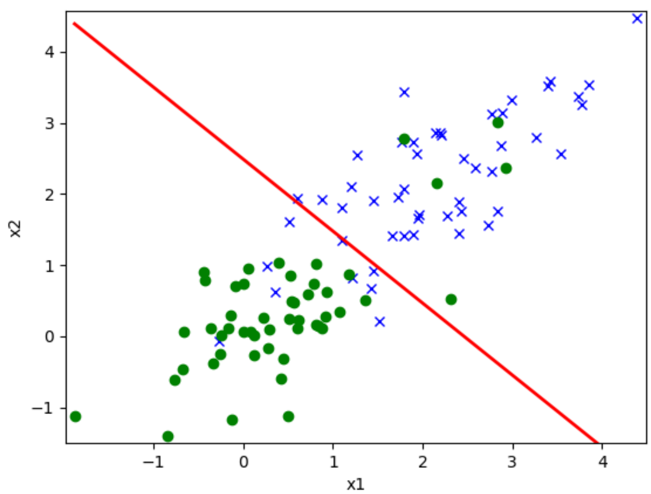

Figure 4: GDA Decision Boundary on Dataset 2

1.6 Question 1 (f)

Comment The decision boundary from logistic regression seems more reasonable. The logistic decision boundary puts weight on both x_1 and x_2. However, the GDA decision boundary is almost vertical and does not put much weight on x_2, the decision is made heavily relied on the $x_1^{(i)}$ for each sample.

1.7 Question 1(g)

Comment Decision boundaries from logistic regression and GDA are nearly identical. In the first dataset GDA seem to perform worse than logistic regression. GDA assumed common variance across groups, however, from the scatter plot, the variance of x_2 within green dot group is much less the variance of x_2 within blue cross group. So the GDA assumption is critically compromised in the first dataset. In contrast, for the second dataset, variance of both x_1 and x_2 are similar in both groups, the common-covariance assumption of GDA model is more realistic, thus GDA and logistic regression generated similar results on the second dataset.

1.8 Question 1(h)

Comment we can normalize/standardize x_1 and x_2 within two groups using

$$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_{x_j}}{\sigma_j} \tag{1.83}$$

so that both x_1 and x_2 have variance of (approximately) 1 in both groups, so that the common-covariance assumption of GDA becomes more realistic.

2 Question 2: Incomplete, Positive-Only Labels

2.1 Question 2(a) Coding

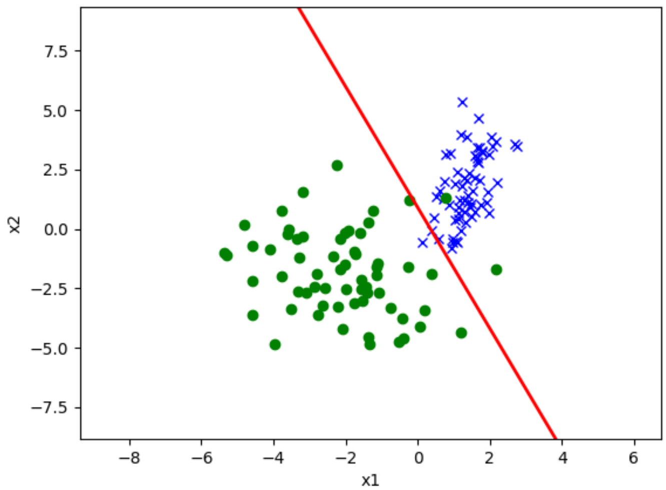

Figure 5: Test Set Prediction of Models Trained on True Label

2.2 Question 2(b) Coding

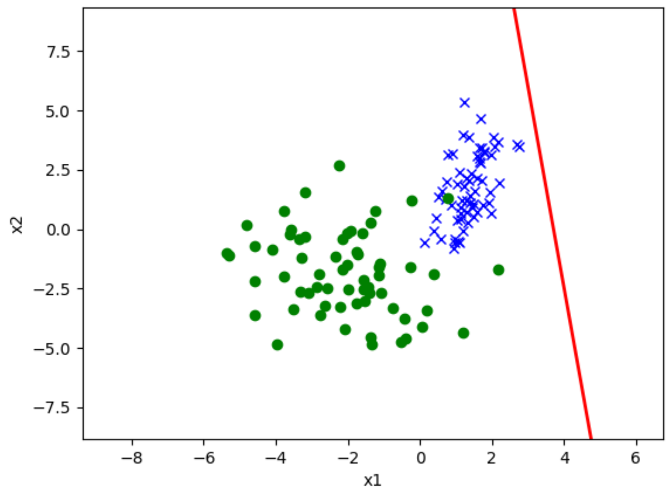

Figure 6: Test Set Prediction of Naive Model

2.3 Question 2(c)

Proof.

$$p(t^{(i)} = 1|y^{(i)} = 1, x^{(i)}) = \frac{p(y^{(i)} = 1|t^{(i)} = 1, x^{(i)})p(t^{(i)} = 1|x^{(i)})}{p(y^{(i)} = 1|x^{(i)})} \tag{2.1}$$

$$= \frac{p(y^{(i)} = 1|t^{(i)} = 1, x^{(i)})p(t^{(i)} = 1|x^{(i)})}{p(y^{(i)} = 1|t^{(i)} = 1, x^{(i)})p(t^{(i)} = 1|x^{(i)}) + p(y^{(i)} = 1|t^{(i)} = 0, x^{(i)})p(t^{(i)} = 0|x^{(i)})} \tag{2.2}$$

$$= \frac{\alpha p(t^{(i)} = 1|x^{(i)})}{\alpha p(t^{(i)} = 1|x^{(i)}) + 0p(t^{(i)} = 0|x^{(i)})} \tag{2.3}$$

$$= \frac{\alpha p(t^{(i)} = 1|x^{(i)})}{\alpha p(t^{(i)} = 1|x^{(i)})} = 1 \tag{2.4}$$

■

2.4 Question 2(d)

Proof.

$$
\begin{aligned}
p(t^{(i)} = 1|x^{(i)}) &= p(t^{(i)}, y^{(i)} = 1|x^{(i)}) + p(t^{(i)} = 1, y^{(i)} = 0|x^{(i)}) && (2.5)\\
&= p(t^{(i)} = 1|y^{(i)} = 1, x^{(i)})p(y^{(i)} = 1|x^{(i)}) && (2.6)\\
&+ p(y^{(i)} = 0|t^{(i)} = 1, x^{(i)})p(t^{(i)} = 1|x^{(i)}) && (2.7)\\
&= 1p(y^{(i)} = 1|x^{(i)}) + (1-\alpha)p(t^{(i)} = 1|x^{(i)}) && (2.8)\\
\implies p(t^{(i)} = 1|x^{(i)}) &= \frac{1}{\alpha}p(y^{(i)} = 1|x^{(i)}) && (2.9)
\end{aligned}
$$

■

2.5 Question 2(e)

Proof.

$$
\begin{aligned}
h(x^{(i)}) &= p(y^{(i)} = 1|x^{(i)}) && (2.10)\\
\implies \mathbb{E}[h(x^{(i)})|y^{(i)} = 1] &= \mathbb{E}[p(y^{(i)} = 1|x^{(i)})|y^{(i)} = 1] && (2.11)\\
&= \mathbb{E}\{p(y^{(i)} = 1|t^{(i)} = 1, x^{(i)})p(t^{(i)} = 1|x^{(i)}) && (2.12)\\
&+ p(y^{(i)} = 1|t^{(i)} = 0, x^{(i)})p(t^{(i)} = 0|x^{(i)})|y^{(i)} = 1\} && (2.13)\\
&= \mathbb{E}[\alpha p(t^{(i)} = 1|x^{(i)}) + 0|y^{(i)} = 1] && (2.14)\\
&= \alpha\mathbb{E}[p(t^{(i)} = 1|x^{(i)})|y^{(i)} = 1] && (2.15)
\end{aligned}
$$

From part (c), we proved that given $y^{(i)} = 1$, $t^{(i)} = 1$ with probability 1, conditioned on $x^{(i)}$. Hence,

$$
\begin{aligned}
\mathbb{E}[p(t^{(i)} = 1|x^{(i)})|y^{(i)} = 1] &= 1 && (2.16)\\
\implies \mathbb{E}[h(x^{(i)})|y^{(i)} = 1] &= \alpha && (2.17)
\end{aligned}
$$

∎

2.6 Question 2(f) Coding

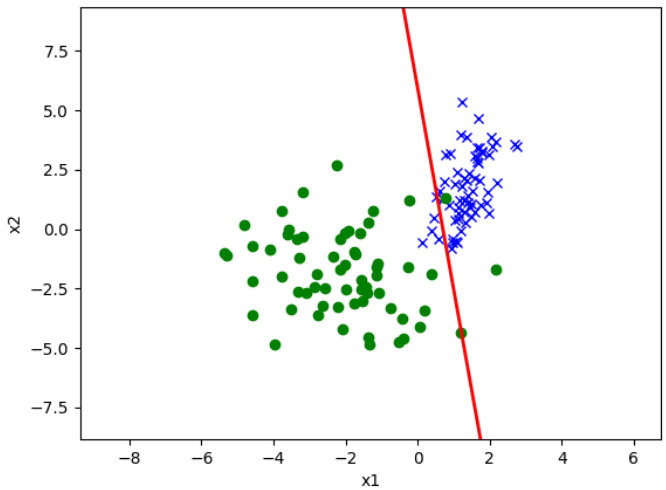

Figure 7: Adjusted Prediction on Test Set

3 Question 3: Poisson Regression

3.1 Question 3(a)

Proof.

$$p(y;\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!} \tag{3.1}$$

$$= \exp\log(\frac{\exp(-\lambda)\lambda^y}{y!}) \tag{3.2}$$

$$= \exp\left(-\lambda + y\log(\lambda) - \log(y!)\right) \tag{3.3}$$

$$= \frac{1}{y!}\exp(\log(\lambda)y - \lambda) \tag{3.4}$$

therefore, Poisson distribution belongs to the exponential family with

$$b(y) := \frac{1}{y!} \tag{3.5}$$

$$\eta(\lambda) := \log(\lambda) \tag{3.6}$$

$$T(y) := y \tag{3.7}$$

$$a(\eta) := \exp(\eta) = \lambda \tag{3.8}$$

■

3.2 Question 3(b)

Answer. By definition, the canonical response function maps η to the expectation $\mathbb{E}[T(y);\eta]$, which equals $\mathbb{E}[y;\eta] = \lambda$ here. Based on the fact that $\eta(\lambda) = \log(\lambda)$, the exponential function maps $\eta(\lambda)$ to $\mathbb{E}[y;\eta]$. Hence, the canonical response function here is the exponential function. ■

3.3 Question 3(c)

Derive.

$$\begin{aligned}
\frac{\partial}{\partial\theta_j}\log(p(y^{(i)}|x^{(i)};\theta)) &= \frac{\partial}{\partial\theta_j}\left(\log(b(y)) + \eta^T y^{(i)} - a(\eta)\right) && (3.9)\\
&= \frac{\partial}{\partial\theta_j}\left(\theta^T x^{(i)} y^{(i)} - \exp(\theta^T x^{(i)})\right) && (3.10)\\
&= x_j^{(i)} y^{(i)} - \exp(\theta^T x^{(i)}) x_j^{(i)} && (3.11)\\
&= \left(y^{(i)} - \exp(\theta^T x^{(i)})\right) x_j^{(i)} && (3.12)
\end{aligned}$$

The stochastic gradient ascent update rule for parameter θ_j is

$$\theta_j \leftarrow \theta_j + \alpha\left(y^{(i)} - \exp(\theta^T x^{(i)})\right) x_j^{(i)} \tag{3.13}$$

where $(x^{(i)}, y^{(i)})$ is the randomly selected sample, and $\alpha > 0$ denotes the learning rate. ■

3.4 Question 3(d) Coding

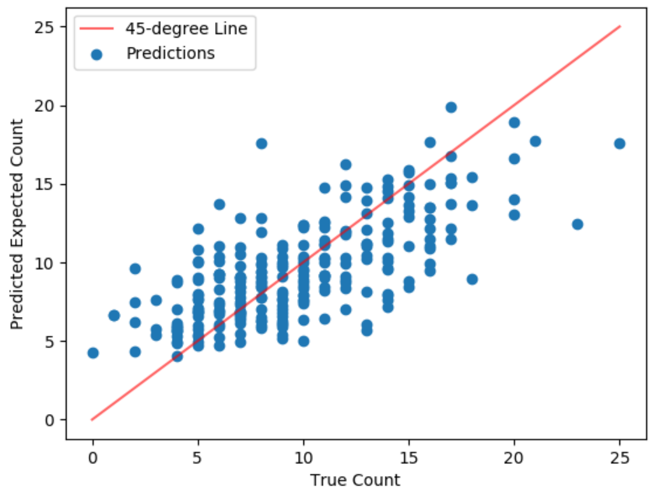

Figure 8: True Count and Predicted Counts on the Validation Set

4 Question 4: Convexity of Generalized Linear Models

4.1 Question 4(a)

Proof. The mean of y is simply

$$\mathbb{E}[Y;\eta] = \int_{\mathbb{R}} yp(y;\eta)\ dy \tag{4.1}$$

$$= \int_{\mathbb{R}} yb(y)\exp(\eta y - a(\eta))\ dy \tag{4.2}$$

By definition of probability measure, it must be the case that

$$\int_{\mathbb{R}} p(y;\eta)\ dy = 1 \tag{4.3}$$

for every valid η. Therefore,

$$\int_{\mathbb{R}} b(y)\exp(\eta y - a(\eta))\ dy = 1 \tag{4.4}$$

$$\implies \int_{\mathbb{R}} b(y)\exp(\eta y)\frac{1}{\exp(a(\eta))}\ dy = 1 \tag{4.5}$$

$$\implies \int_{\mathbb{R}} b(y)\exp(\eta y)\ dy = \exp(a(\eta)) \tag{4.6}$$

$$\implies \frac{\partial \exp(a(\eta))}{\partial \eta} = \frac{\partial}{\partial \eta}\int_{\mathbb{R}} b(y)\exp(\eta y)\ dy \tag{4.7}$$

$$\implies a'(\eta)\exp(a(\eta)) = \int_{\mathbb{R}} b(y)\frac{\partial \exp(\eta y)}{\partial \eta}\ dy \tag{4.8}$$

$$\implies a'(\eta) = \int_{\mathbb{R}} yb(y)\exp(\eta y)\frac{1}{\exp(a(\eta))}\ dy \tag{4.9}$$

$$\implies a'(\eta) = \int_{\mathbb{R}} yb(y)\exp(\eta y - a(\eta))\ dy \tag{4.10}$$

$$= \mathbb{E}[Y;\eta] \tag{4.11}$$

■

4.2 Question 4(b)

Proof. From part (a),

$$
\begin{aligned}
a'(\eta) &= \int_{\mathbb{R}} yb(y)\exp(\eta y - a(\eta))\ dy && (4.12)\\
\implies \frac{\partial^2 a(\eta)}{\partial \eta^2} &= \frac{\partial}{\partial \eta}\int_{\mathbb{R}} yb(y)\exp(\eta y - a(\eta))\ dy && (4.13)\\
&= \int_{\mathbb{R}} yb(y)\exp(ny - a(\eta))\big(y - a'(\eta)\big)\ dy && (4.14)\\
&= \int_{\mathbb{R}} y^2 b(y)\exp(ny - a(\eta))\ dy - a'(\eta)\int_{\mathbb{R}} yb(y)\exp(ny - a(\eta))\ dy && (4.15)\\
&= \mathbb{E}[Y^2;\eta] - a'(\eta)\mathbb{E}[Y;\eta] && (4.16)\\
&= \mathbb{E}[Y^2;\eta] - \mathbb{E}[Y;\eta]^2 && (4.17)\\
&= \mathbb{V}[Y;\eta] && (4.18)
\end{aligned}
$$

■

4.3 Question 4(c)

Proof. The negative log-likelihood (NLL) loss function looks like

$$\ell(\theta) = -\log\left(\prod_{i=1}^{n} p(y^{(i)}; \eta^{(i)})\right) \tag{4.19}$$

$$= -\log\left(\prod_{i=1}^{n} p(y^{(i)}; \theta^T x^{(i)})\right) \tag{4.20}$$

$$= -\sum_{i=1}^{n} \log(p(y^{(i)}; \theta^T x^{(i)})) \tag{4.21}$$

$$= -\sum_{i=1}^{n} \log(b(y^{(i)})) + [\theta^T x^{(i)} - a(\theta^T x^{(i)})] \tag{4.22}$$

Then for each $j \in \{1, \ldots, d\}$, the first order derivative of $\ell(\theta)$ becomes

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = -\sum_{i=1}^{n} x_j^{(i)} - \frac{\partial a(\theta^T x^{(i)})}{\partial \eta} \frac{\partial \eta}{\partial \theta_j} \tag{4.23}$$

$$= -\sum_{i=1}^{n} x_j^{(i)} - \mathbb{E}[Y; \eta] x_j^{(i)} \tag{4.24}$$

then for each $k \in \{1, \ldots, d\}$, the second order derivative is

$$\frac{\partial^2 \ell(\theta)}{\partial \theta_j \theta_k} = -\sum_{i=1}^{n} -x_j^{(i)} \frac{\partial \mathbb{E}[Y; \eta]}{\partial \eta} \frac{\partial \eta}{\partial \theta_k} \tag{4.25}$$

$$= \sum_{i=1}^{n} x_j^{(i)} x_k^{(i)} \mathbb{V}[Y; \eta] \tag{4.26}$$

Let $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$ be an arbitrary vector in $\mathbb{R}^d$,

$$z^T H_\ell(\theta) z = \sum_{k=1}^{d} \sum_{j=1}^{d} \sum_{i=1}^{n} z_k x_k^{(i)} x_j^{(i)} z_j \mathbb{V}[Y; \eta] \tag{4.27}$$

$$= \mathbb{V}[Y; \eta] \sum_{i=1}^{n} \sum_{k=1}^{d} \sum_{j=1}^{d} z_k x_k^{(i)} x_j^{(i)} z_j \tag{4.28}$$

$$= \mathbb{V}[Y; \eta] \sum_{i=1}^{n} (x^{(i)} z)^2 \text{ (by hint in Q1.a)} \tag{4.29}$$

Further, the variance of random variable is always nonnegative, and $(x^{(i)} z)^2 \geq 0$. Therefore, for every $z \in \mathbb{R}^d$,

$$z^T H_\ell(\theta) z \geq 0 \tag{4.30}$$

$H_\ell(\theta)$ is positive semidefinite. $\blacksquare$

5 Question 5: Linear Regression

5.1 Question 5(a)

$$J(\theta) := \frac{1}{2}\sum_{i=1}^{n}\left(y^{(i)} - \theta^T\phi(x^{(i)})\right)^2 \tag{5.1}$$

$$\theta \leftarrow \theta + \alpha\sum_{i=1}^{n}\left(y^{(i)} - \theta^T\hat{x}^{(i)}\right)\hat{x}^{(i)} \tag{5.2}$$

where α denotes the learning rate.

5.2 Question 5(b) Coding

Comment It is theoretically possible to fit a sine function using polynomials by Taylor's expansion. However, when $k = 3$, the polynomial model is not expressive enough to capture all curvatures in the training set.

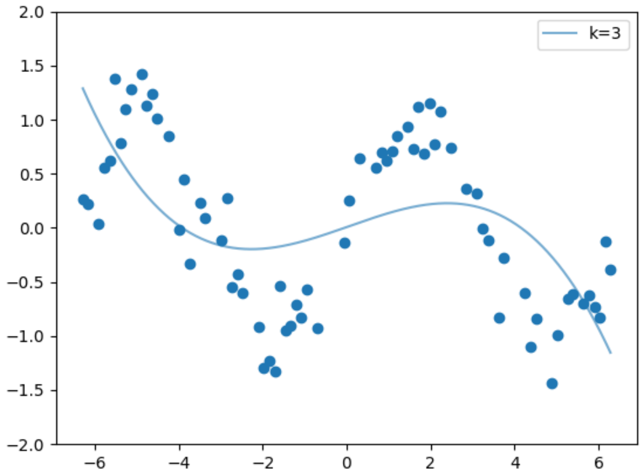

Figure 9: Polynomial Features up to Degree 3

5.3 Question 5(c) Coding

Comment For models with $k = 5$ or $k = 10$, the model captures the curvature of training samples much better than $k = 3$ model does. However, for $k = 20$, the model starts to over fit the noise in the training set, and the expected value demonstrates abnormal curvature near $x = -3$, $x = 4.5$, and $x = 6$.

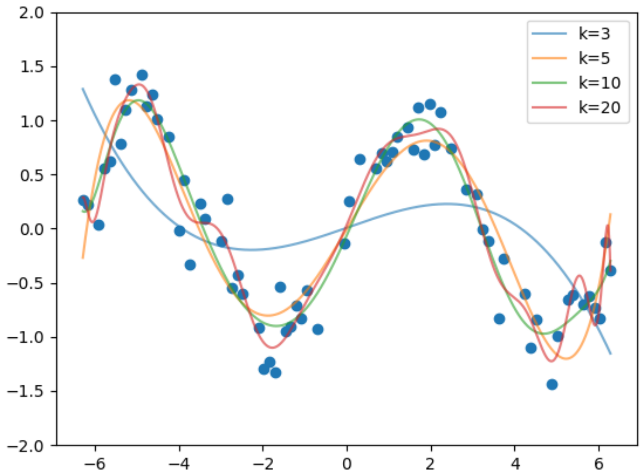

Figure 10: Polynomial Features

5.4 Question 5(d) Coding

Comment By observing the training set, it can be postulated that the true data generating process actually involves a sin function. Adding the $\sin(x)$ feature allows the model to fully capture the underlying pattern, even with low degree of polynomial terms. Also, when k is large, the model starts to over fit training samples by capturing the noise.

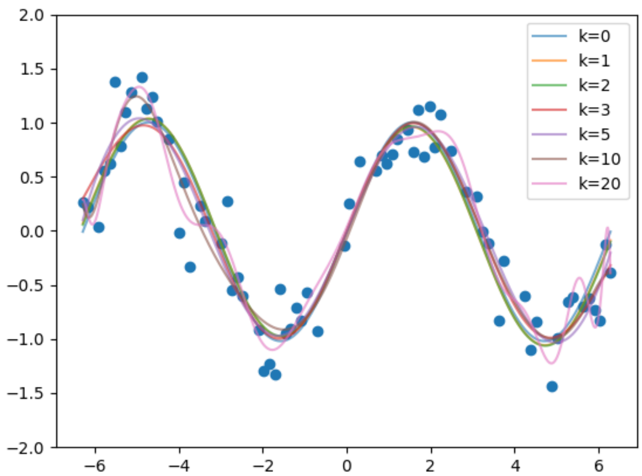

Figure 11: Polynomial and Sin Features

5.5 Question 5(e) Coding

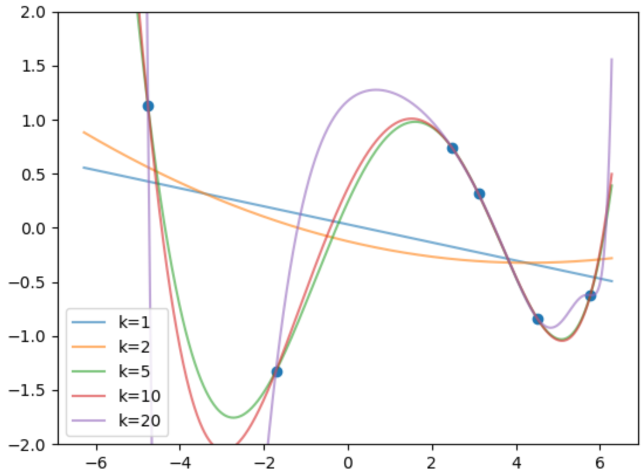

Figure 12: Overfitting on Small Dataset

Comment When the number of features is smaller than number of training samples, the model is experiencing under fitting problem. However, when the number of parameters in the model is more than the number of training, the model can fit all training samples exactly (when $k \geq 5$, the fitting line passes through all training samples precisely), and the model suffers from over fitting problem.