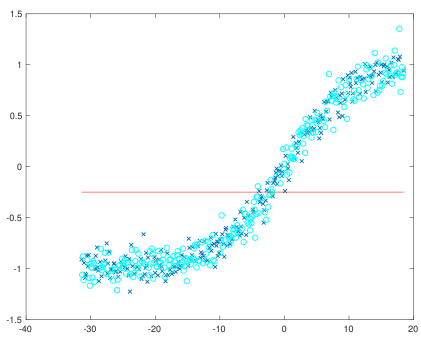
Homework 1

Yunian Pan

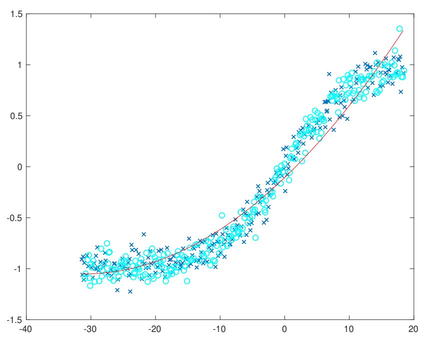
February 10, 2019

1 Problem 1: Overfitting

Select some typical values of d from {1, ..., 20}, the fitting lines are shown as below:



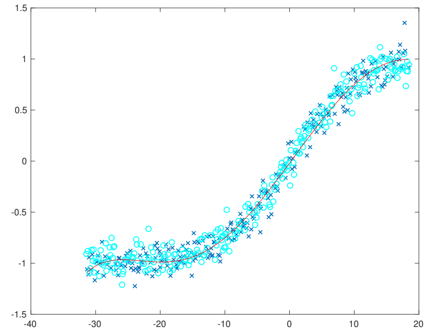
(a) d=1



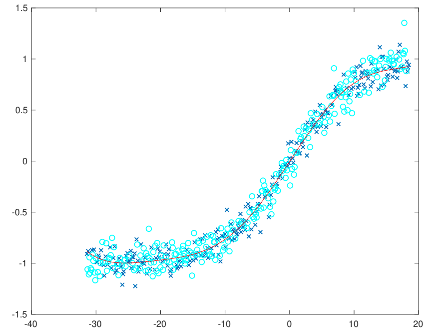
(b) d=3

Obviously the functions from 1(c) and 1(d) best fit the data set. Through cross-validation, we can get the best $d = 9$, which corresponds with the lowest loss on testing set, as shown in 1(g) and 1(h).

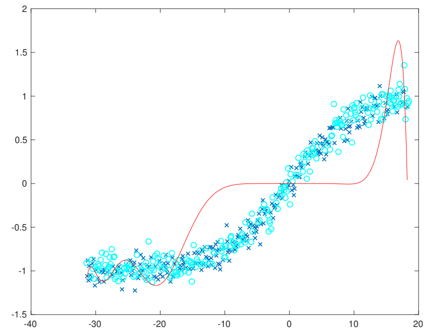
The reason why training error keeps rising after $d = 10$ may be as we increase the polynomial degree, the disturbance between training examples becomes too large so that the coefficients explodes, making the training model unfit the training data.



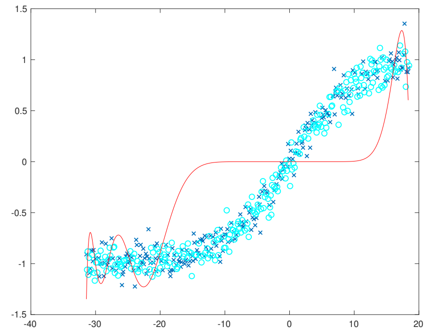
(c) d=6



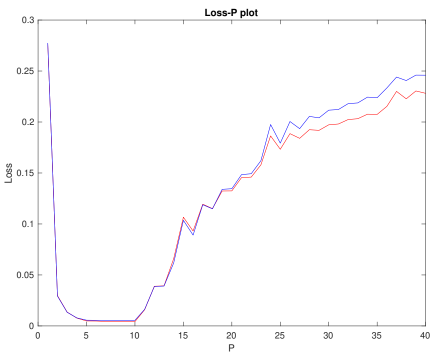
(d) d=9



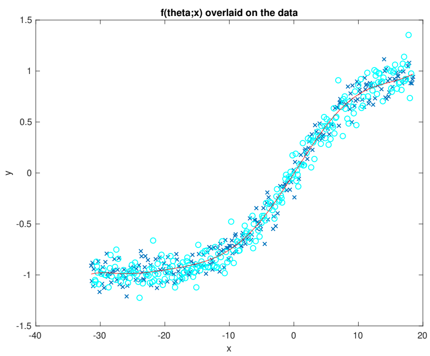
(e) d=15



(f) d=20



(g) cross-validation



(h) overlaid

2 Problem 2: logistic Regression

For the function:

$$f(\mathbf{x};\theta) = (1 + \exp(-\theta^\intercal \mathbf{x}))^{-1}$$

with the loss function:

$$L = (1 - y_i)\log(1 - f(x_i, \theta)) - y_i \log(f(x_i, \theta))$$

We were supposed to solve the gradient descent with derivation: $\nabla_\theta L = 0$

$$\begin{aligned}
\nabla_\theta L &= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(1-y_i)}{1-f(x_i;\theta)} - \frac{y_i}{f(x_i;\theta)}\right] f^{'}(x_i;\theta) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(1-y_i)(1+e^{-\theta^\intercal x_i})}{e^{-\theta^\intercal x_i}} - y_i(1+e^{-\theta^\intercal x_i})\right]\frac{x_i e^{-\theta^\intercal x_i}}{(1+e^{-\theta^\intercal x_i})^2} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(1-y_i)x_i}{1+e^{-\theta^\intercal x_i}} - \frac{y_i x_i e^{-\theta^\intercal x_i}}{1+e^{-\theta^\intercal x_i}}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[(1-y_i)x_i - y_i x_i e^{-\theta^\intercal x_i}\right](1+e^{-\theta^\intercal x_i})^{-1} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[x_i - y_i x_i - y_i x_i(f(x_i;\theta)^{-1} - 1)\right] f(x_i;\theta) \\
&= \frac{1}{N}\sum_{i=1}^{N}(x_i f(x_i;\theta) - y_i x_i)
\end{aligned}$$

However, this equation can't be solved analytically, only with recursive numerical method can we approach to the convex point.

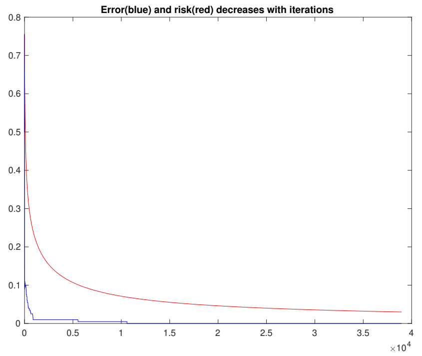
Using GD, set the iteration to be 200000 with tolerance $\epsilon = 0.001$, step size $\eta = 2$, the model will be as shown as $1(i)$ and $1(j)$.

The model obtained is $\theta = [-26.5723, -117.8034, 9.0470]^\top$. Slightly change the tolerance as well as the step size, there's no notable difference regarding the convergence evolution rate and the model accuracy.

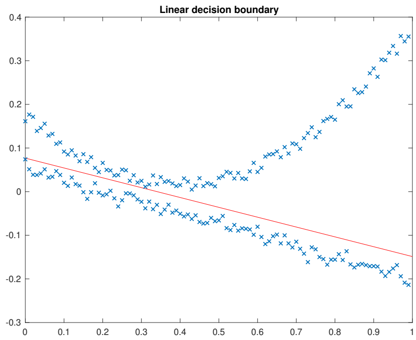
Using SGD, set the batch size $b = 1$, step size $\eta = 2$, the model will be as shown as $1(k)$ and $1(l)$.

The model obtained is $\theta = [-11.9161, -58.4613, 4.3810]^\top$, which is quite the same as the previous one. increasing the batch size, the convergence rate will slow down a little, yet greater than GD.

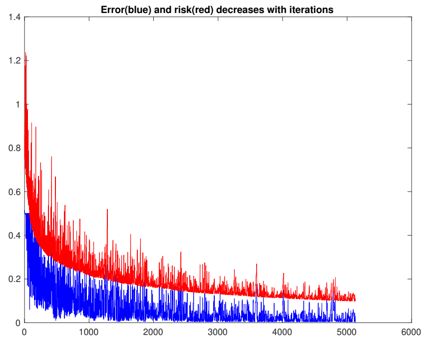
Thus we conclude that SGD is more unstable while it converges faster than GD.



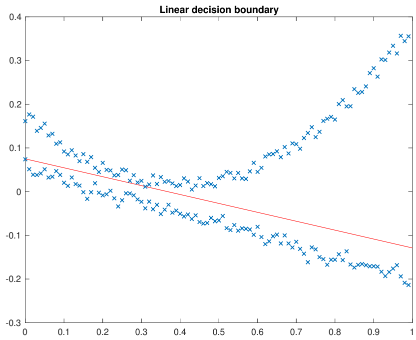
(i) error and risk



(j) binary boundary



(k) error and risk



(l) binary boundary

3 problem 3: Multi-class Discrimination

3.1 a)

In the 2-dimensional case, $K = 3$, suppose there are 2 functions $y_1(x)$ and $y_2(x)$, according to the rule, when $y_1(x) > 0$ && $y_2(x) > 0$, the class of x can be both C_1 or C_2, so it can't be determined, when $y_1(x) < 0$ && $y_2(x) < 0$, we only know x belongs to none of the 2 classes. Thus the approach leads to the ambiguous region x-space shown in 3.1 denoted by the orange area. Unless $y_1(x)$ and $y_2(x)$ are the same, otherwise there's always an ambiguous region

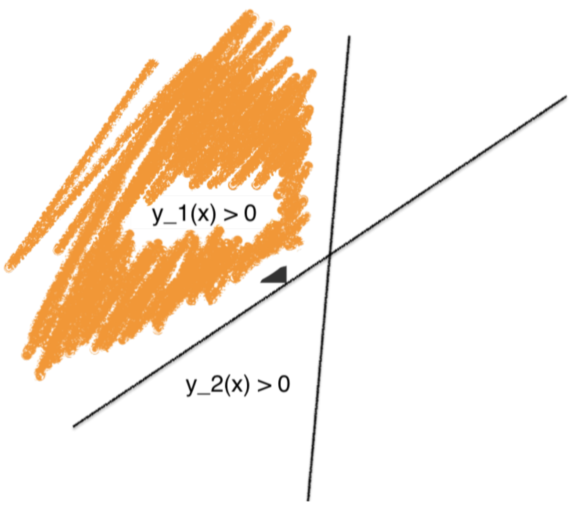


Figure 3.1

3.2 b)

According to the discrimination strategy, there are $K(K-1)/2 = 3$ linear functions when $k = 3$, which can be defined as follows:

$$
\begin{array}{llll}
y_{12}(x) > 0 & x \in C_1 & y_{12}(x) < 0 & x \in C_2 \\
y_{13}(x) > 0 & x \in C_1 & y_{12}(x) < 0 & x \in C_3 \\
y_{23}(x) > 0 & x \in C_2 & y_{12}(x) < 0 & x \in C_3
\end{array}
$$

the structure is not well-defined because of following paradox:

$$
\begin{array}{ll}
y_{12}(x) > 0 \ \&\& \ y_{13}(x) < 0 & x \in C_1 \ \&\& \ x \in C_3 \\
y_{13}(x) > 0 \ \&\& \ y_{23}(x) < 0 & x \in C_1 \ \&\& \ x \in C_2 \\
y_{23}(x) > 0 \ \&\& \ y_{13}(x) < 0 & x \in C_2 \ \&\& \ x \in C_3 \\
y_{12}(x) > 0 \ \&\& \ y_{13}(x) < 0 \ \&\& \ y_{23}(x) > 0 & x \in C_1 \ \&\& \ x \in C_3 \ \&\& \ x \in C_2
\end{array}
$$

Which leads to the ambiguous region in 3.2:

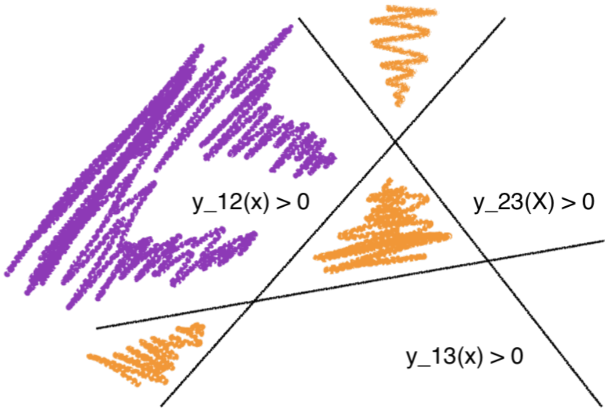


Figure 3.2