Problem Set 1, CS229(Machine Learning)

Ma Yubo

June 20th, 2021
(Modified on July, 11th, 2021)

1 Linear Classifiers (Logistic Regression and GDA)

1.1

Let μ be $h_\theta(x) = 1/(1+e^{-\theta^T x})$. The gradient about θ of loss function J is:

$$\begin{aligned}
&\nabla_\theta L(\theta) = \nabla_\mu L(\theta) \nabla_\theta \mu \\
&= -\frac{1}{m}\sum_{i=1}^{m}(y^{(i)}\frac{1}{h_\theta(x^{(i)})}\nabla_\theta \mu + (1-y^{(i)})\frac{-1}{1-h_\theta(x^{(i)})}\nabla_\theta \mu) \\
&= -\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - h(x^{(i)}))x^{(i)}
\end{aligned} \tag{1.1}$$

Since $\nabla_\theta \mu = (1-h_\theta(x))h_\theta(x)$. For each component j

$$\frac{\partial L}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - h(x^{(i)}))x_j^{(i)} \tag{1.2}$$

$$\begin{aligned}
&\frac{\partial^2 L}{\partial \theta_j \theta_k} \\
&= \frac{1}{m}\sum_{i=1}^{m}\frac{\partial h(x^{(i)})}{\partial \theta_k}x_j^{(i)} \\
&= \frac{1}{m}\sum_{i=1}^{m} h(x^{(i)})(1-h(x^{(i)})){x_k^{(i)}}^T x_j^{(i)}
\end{aligned} \tag{1.3}$$

Accoding to the definition of Hessian matrix, $\frac{\partial^2 L}{\partial \theta_j \theta_k}$ is the element (row j, column k) of $H(\theta)$. Thus,

$$\begin{aligned}
z^T H z &= \frac{1}{m}\sum_{i=1}^{m} h(x^{(i)})(1-h(x^{(i)}))\sum_{j=1}^{n}\sum_{k=1}^{n} z_j z_k {x_k^{(i)}}^T x_j^{(i)} \\
&= \frac{1}{m}\sum_{i=1}^{m} h(x^{(i)})(1-h(x^{(i)}))(\sum_{j=1}^{n} z_j x_j^{(i)})^2 \\
&\geq 0
\end{aligned} \tag{1.4}$$

since $h(x^{(i)}) \in (0,1), \forall i \in \{1,2,...,m\}$. By the definition of positive matrix, we know that $z^T H z \geq 0$ holds true.

1.2

Coding Problem. See related source code.

1.3

By Beyesian formula,

$$\begin{aligned}
&P(y=1|x;\Theta) \\
&= \frac{P(x|y=1;\mu_1,\Sigma)P(y=1;\phi)}{P(x|y=1;\mu_1,\Sigma)P(y=1;\phi)+P(x|y=0;\mu_0,\Sigma)P(y=0;\phi)} \\
&= \frac{\phi exp\left\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right\}}{\phi exp\left\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right\}+(1-\phi)exp\left\{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right\}} \\
&= \frac{1}{1+\frac{1-\phi}{\phi}exp\left\{\frac{1}{2}[(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)-(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]\right\}}
\end{aligned} \tag{1.5}$$

Let $\theta = \Sigma^{-1}(\mu_1 - \mu_0)$, $\theta_0 = \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1) + log(\frac{\phi}{1-\phi})$. Then the posterior distribution can be written as the format of logistic regression:

$$p(y=1|x;\Theta) = \frac{1}{1+exp(-(\theta_0+\theta^T x))} \tag{1.6}$$

1.4

The log-likelihood of the data is:

$$l(\Theta) = log\prod_{i=1}^{m} P(x^{(i)}, y^{(i)}; \Theta) \tag{1.7}$$

By maximizing $l(\phi)$ we have:

$$
\begin{aligned}
&argmax_{\Theta} l(\Theta) = argmax_{\Theta} log \prod_{i=1}^{m} P(x^{(i)}, y^{(i)}; \Theta) \\
&= argmax_{\Theta} log \prod_{i=1}^{m} P(x^{(i)}|y^{(i)}; \Theta) P(y^{(i)}; \Theta) \\
&= argmax_{\Theta} \sum_{i=1}^{m} [(1-y) log(P(x|y=0; \mu_0, \Sigma) + y log(P(x|y=1; \mu_1, \Sigma) + (1-y) log(1-\phi) + y log(\phi)] \\
&= argmax_{\Theta} \sum_{i=1}^{m} [\frac{-1}{2}(1-y)(x-\mu_0)^T \Sigma^{-1}(x-\mu_0) + \frac{-1}{2} y (x-\mu_1)^T \Sigma^{-1}(x-\mu_1) + \\
&(1-y) log(1-\phi) + y log(\phi) - log(|\Sigma|]
\end{aligned} \tag{1.8}
$$

Here $\Theta = (\phi, \mu_0, \mu_1, \Sigma)$. We take derivation of the equation above on these component respectively.

$$\frac{\partial L}{\partial \phi} = \sum_{i=1}^{m} \frac{y^{(i)} - 1}{1-\phi} + \frac{y^{(i)}}{\phi} = 0 \tag{1.9}$$

$$\Rightarrow \phi = \frac{1}{m} \sum_{i=1}^{m} I(y^{(i)} = 1) \tag{1.10}$$

$$\frac{\partial L}{\partial \mu_0} = \Sigma^{-1} \sum_{i=1}^{m} (y^{(i)} - 1)(\mu_0 - x^{(i)}) = 0 \tag{1.11}$$

$$\Rightarrow \mu_0 = \frac{1}{m} \frac{\sum_{i=0}^{m} I(y^{(i)} = 0) x^{(i)}}{\sum_{i=0}^{m} I(y^{(i)} = 0)} \tag{1.12}$$

$$\frac{\partial L}{\partial \mu_1} = \Sigma^{-1} \sum_{i=1}^{m} (y^{(i)} - 1)(\mu_1 - x^{(i)}) = 0 \tag{1.13}$$

$$\Rightarrow \mu_1 = \frac{1}{m} \frac{\sum_{i=0}^{m} I(y^{(i)} = 1) x^{(i)}}{\sum_{i=0}^{m} I(y^{(i)} = 1)} \tag{1.14}$$

The derivation of Σ is untrivial. For simplicity, we assume that $\mu_0 = \mu_1 = \mu$.

$$L(\Sigma) = -\frac{m}{2} log(|\Sigma|) - \sum_{i=1}^{m} (x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu) \tag{1.15}$$

Then,

$$
\begin{aligned}
dL(\Sigma) &= d(trL(\Sigma)) = tr[-\frac{m}{2|\Sigma|}d|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(x^{(i)}-\mu)^T d\Sigma^{-1}(x^{(i)}-\mu)] \\
&= tr[-\frac{m}{2|\Sigma|}|\Sigma|tr(\Sigma^{-1}d\Sigma) - \frac{1}{2}\sum_{i=1}^{m}(x^{(i)}-\mu)^T(-\Sigma^{-1}d\Sigma\Sigma^{-1})(x^{(i)}-\mu)] \\
&= tr[\frac{1}{2}(-m\Sigma^{-1} + \Sigma^{-1}[\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T]\Sigma^{-1})d\Sigma]
\end{aligned}
\tag{1.16}
$$

By $dL = tr(\frac{\partial L}{\partial \Sigma}d\Sigma)$, we know that:

$$
\frac{\partial L}{\partial \Sigma} = \frac{1}{2}P^{-1}(m\Sigma^{-1} - \Sigma^{-1}[\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T]\Sigma^{-1})P \tag{1.17}
$$

where P is an invertible matrix. To maximize L, take $\frac{\partial L}{\partial \Sigma}$ as 0:

$$
\Rightarrow -m\Sigma^{-1} + \Sigma^{-1}[\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T]\Sigma^{-1} = 0 \tag{1.18}
$$

$$
\Rightarrow -m\Sigma^{-1} + \Sigma^{-1}[\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T]\Sigma^{-1} = 0 \tag{1.19}
$$

$$
\Rightarrow \Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T \tag{1.20}
$$

Now, with two groups within which belongs to different gaussian distributions(mean), we have:

$$
\begin{aligned}
\Sigma &= \frac{1}{m_1}\sum_{y^{(i)}=1}(x^{(i)}-\mu_1)(x^{(i)}-\mu_1)^T + \frac{1}{m_0}\sum_{y^{(i)}=0}(x^{(i)}-\mu_0)(x^{(i)}-\mu_0)^T \\
&= \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T
\end{aligned}
\tag{1.21}
$$

1.5

Coding Problem. See related source code.

1.6

The samples of validation set in dataset 1 and the decision boundaries from logistic regression and GDA are plotted below.

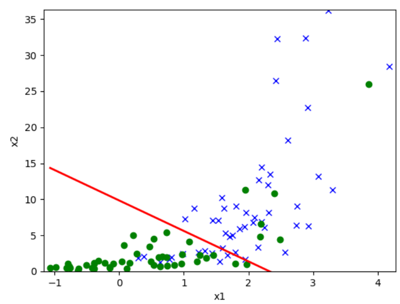

Figure 1.1: Scatter plot of dataset 1 and decision boundary for LR

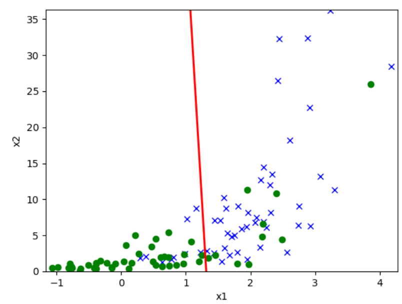

Figure 1.2: Scatter plot of dataset 1 and decision boundary for GDA

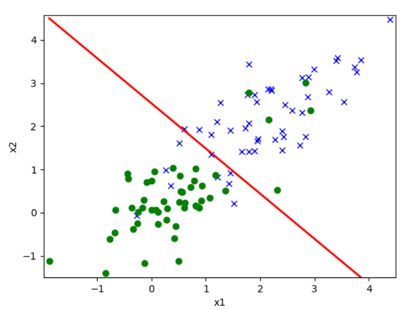

Figure 1.3: Scatter plot of dataset 2 and decision boundary for LR

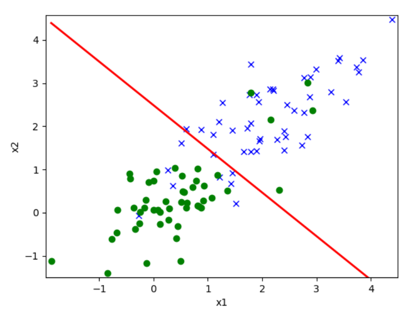

Figure 1.4: Scatter plot of dataset 2 and decision boundary for GDA

1.7

The samples of validation set in dataset 1 and the decision boundaries from logistic regression and GDA are plotted above. GDA performs worse on dataset 2 compared with dataset 1. Since GDA holds a stronger assumption that the data is generated by **Gaussian** distributions, it performs not well on non-gaussian data (such as data in set 1).

1.8

Observe that all data's $x^{(2)}$ are non-negative in dataset 1. So we can take log-transformation on $x^{(2)}$ axis.

2 Incomplete, Positive-Only Labels

2.1

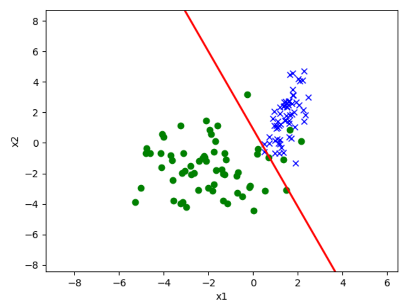

Figure 2.1: fully observed case

2.2

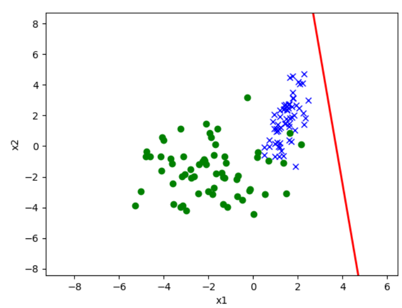

Figure 2.2: naive method on partial labels

2.3

By bayesian formula,

$$
\begin{aligned}
&P(t^{(i)} = 1|y^{(i)} = 1, x^{(i)}) \\
&= \frac{P(y^{(i)} = 1|t^{(i)} = 1, x^{(i)})P(t^{(i)} = 1|x^{(i)})}{P(y^{(i)} = 1|t^{(i)} = 1, x^{(i)})P(t^{(i)} = 1|x^{(i)}) + P(y^{(i)} = 1|t^{(i)} = 0, x^{(i)})P(t^{(i)} = 0|x^{(i)})} \\
&= \frac{1}{1+0} = 1
\end{aligned}
\tag{2.1}
$$

2.4

For points with true labels,

$$P(t^{(i)} = 1|X^{(i)}) = 1 \tag{2.2}$$

$$\begin{aligned} &P(y^{(i)} = 1|X^{(i)}) \\ &= P(y^{(i)} = 1|t^{(i)} = 1, x^{(i)}) + P(y^{(i)} = 1|t^{(i)} = 0, x^{(i)}) \\ &= \alpha + 0 = \alpha \end{aligned} \tag{2.3}$$

For points with false labels,

$$P(t^{(i)} = 1|X^{(i)}) = 0, P(y^{(i)} = 1|X^{(i)}) = 0, \tag{2.4}$$

In summary, we have $P(t^{(i)} = 1|X^{(i)} = \frac{1}{\alpha} P(y^{(i)} = 1|X^{(i)}$

2.5

Assume that we have magically obtained a function h(x) that perfectly predicts the value of $p(y^{(i)} = 1|x^{(i)})$. That is, $h(x^{(i)}) = p(y^{(i)} = 1|x^{(i)})$.

- If $y^{(i)} = 1$, we have $h(x^{(i)}) = p(y^{(i)} = 1|x^{(i)})$ as stated above, then, $h(x^{(i)}) = \alpha$.
- If $y^{(i)} = 0$, we still have $h(x^{(i)}) = p(y^{(i)} = 1|x^{(i)})$ as stated above, then $h(x^{(i)}) = 0$.

Therefore, $\alpha = E[h(x^{(i)})|y^{(i)} = 1]$

2.6

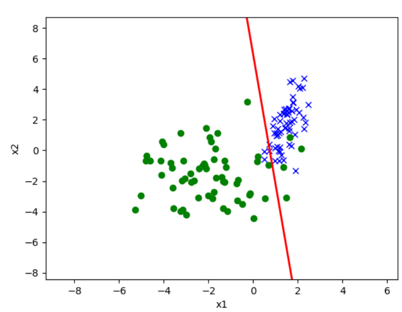

Figure 2.3: naive method on adjusted labels

3 Poisson Regression

3.1

The definition of exponential family distribution:

$$p(y;\eta) = b(y)exp\left\{\eta^T T(y) - a(\eta)\right\} \tag{3.1}$$

We can re-write the p.d.f. of poisson distribution:

$$\begin{aligned} p(y;\eta) &= \frac{\lambda^y}{y!}exp\{-\lambda\} \\ &= \frac{1}{y!}exp\{-\lambda + ylog(\lambda)\} \end{aligned} \tag{3.2}$$

Then $b(y) = 1/y!$, $T(y) = y$, $\eta = log(\lambda)$, $a(\eta) = \lambda = exp(\eta)$.

3.2

Our goal is to predict the expected value of y given x, which means we would like the canonical response function $h(x)$ satisfying $h(x) = E[y|x]$. Therefore,

$$h(x) = E[y|x] = \lambda = exp(\eta) \tag{3.3}$$

3.3

The natural parameter η and the inputs x are related linearly: $\eta = \theta^T x$. So we have:

$$\begin{aligned} P(y|x;\theta) &= \frac{1}{y!}exp\{-\lambda + ylog(\lambda)\} \\ &= \frac{1}{y!}exp\left\{-exp(\theta^T x) + y\theta^T x\right\} \end{aligned} \tag{3.4}$$

The NLL function:

$$logP = \sum_{i=1}^{m} exp\left\{\theta^T x^{(i)}\right\} - y^{(i)}\theta^T x^{(i)} + log(y^{(i)}) \tag{3.5}$$

Maximizing it:

$$\frac{\partial logP}{\partial\theta} = \sum_{i=1}^{m}[exp\left\{\theta^T x^{(i)}\right\} x^{(i)} - y^{(i)}x^{(i)}] \tag{3.6}$$

So the updating rule is:

$$\theta = \theta - \alpha\sum_{i=1}^{m}[exp\left\{\theta^T x^{(i)}\right\} - y^{(i)}]x^{(i)} \tag{3.7}$$

3.4

Apply poisson regression on dataset provided. Draw scatte plot about true count and predition count by poisson regression.

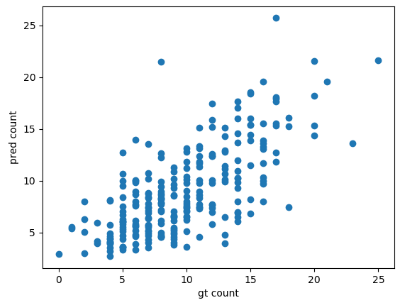

Figure 3.1: true count v.s pred count

4 Convexity of Generalized Linear Models

4.1

By the definition of exponential distribution family, we have

$$exp\left\{a(\eta)\right\} = b(y)exp\left\{\eta y\right\} \tag{4.1}$$

Take derivation on η:

$$exp\left\{a(\eta)\right\}a^{'}(\eta) = b(y)exp\left\{\eta y\right\}y \tag{4.2}$$

$$\Rightarrow a^{'}(\eta)\int 1dy = \int y[b(y)exp\left\{\eta y - a(\eta)\right\}]dy \tag{4.3}$$

$$\Rightarrow a^{'}(\eta) = E[y] \tag{4.4}$$

4.2

Similarly,

$$(a^{''}(\eta) + a^{'}(\eta)^2)exp\left\{a(\eta)\right\} = b(y)exp\left\{\eta y\right\}y^2 \tag{4.5}$$

$$\Rightarrow a^{''}(\eta)\int 1dy = \int y^2[b(y)exp\left\{\eta y - a(\eta)\right\}]dy - a^{'}(\eta)^2\int 1dy \tag{4.6}$$

$$\Rightarrow a^{''}(\eta) = E[y^2] - E^2[y] = Var[y] \tag{4.7}$$

4.3

Write out the NLL function:

$$l(\eta) = -\sum_{i=1}^{m} logP(y^{(i)}|x^{(i)}) = -\sum_{i=1}^{m} a(\theta^T x^{(i)}) - \theta^T x^{(i)}y^{(i)} - log(b(y^{(i)})) \tag{4.8}$$

$$\Rightarrow \frac{\partial^2 l(\theta)}{\partial\theta_j \partial\theta_k} = \sum_{i=1}^{m} a^{''}(\theta^T x^{(i)}) x_j^{(i)} x_k^{(i)} \tag{4.9}$$

$$\Rightarrow \nabla_\theta l(\theta) = \sum_{i=1}^{m} a^{''}(\theta^T x^{(i)}) x^{(i)} {x^{(i)}}^T \tag{4.10}$$

By proved above, $a''(\theta x^{(i)}) = Var[y^{(i)}] \geq 0$. Thus

$$\begin{aligned}
&\Rightarrow z^T \nabla_\theta z l(\theta) = z^T \sum_{i=1}^{m} a^{''}(\theta^T x^{(i)}) x^{(i)} {x^{(i)}}^T z \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{n} a^{''}(\theta^T x^{(i)})(z_j x_j^{(i)})(z_k x_k^{(i)}) \qquad (4.11) \\
&= \sum_{i=1}^{m} a^{''}(\theta^T x^{(i)}) (\sum_{j=1}^{n} z_j x_j^{(i)})^2 \geq 0
\end{aligned}$$

So $\nabla_\theta l(\theta)$ is PSD matrix.

5 Locally Weighted Linear Regression

5.1

5.1.1

By the definition of matrix quadratic form, we have:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)} - y^{(i)})^2 = (X\theta - Y)^T W (X\theta - Y) \tag{5.1}$$

where W = $\begin{bmatrix} w^{(1)} & \dots & \dots & \dots \\ \dots & w^{(2)} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & w^{(m)} \end{bmatrix}$

5.1.2

$$\nabla J(\theta) = 2X^T W X\theta - 2X^T W y = 0 \tag{5.2}$$

$$\Rightarrow X^T W X\theta = X^T W y \tag{5.3}$$

$$\Rightarrow \theta = (X^T W X)^{-1} X^T W y \tag{5.4}$$

5.1.3

Write out the NLL function:

$$L = \sum_{i=1}^{m} -log(\sigma^{(i)} - \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2} \tag{5.5}$$

Since $\sigma^{(i)}$'s are constant, we have:

$$\begin{aligned} argminL &= argmin \sum_{i=1}^{m} -\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2} \\ &= argmin(Y - X\theta)^T W (Y - X\theta) \end{aligned} \tag{5.6}$$

where $W = \begin{bmatrix} 1/(\sigma^{(1)})^2 & \dots & \dots & \dots \\ \dots & 1/(\sigma^{(2)})^2 & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ \dots & \dots & \dots & 1/(\sigma^{(m)})^2 \end{bmatrix}$

5.2

The results of model with $\tau = 0.5$ are plottd below: The MSE value on validation set is 0.293 and the data seems to be under-fitting.

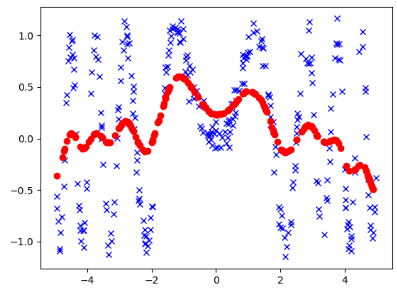

Figure 5.1: Prediction result ($\tau = 0.5$)

5.3

We will now tune the hyperparameter τ. The MLE results are shown as below. The best τ on validation set is 0.1. And the model with this tau gets 0.167 MSE

Table 5.1: MSE results of different τ's

τ	MSE
0.03	0.369
0.05	0.038
0.1	0.136
0.5	0.293
1.0	0.396
10.0	0.438

value on test set.

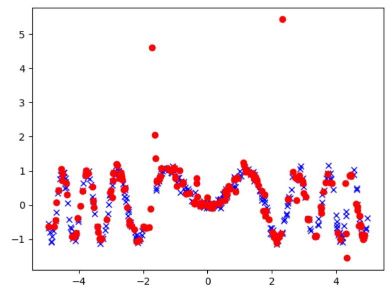

Figure 5.2: $\tau = 0.03$

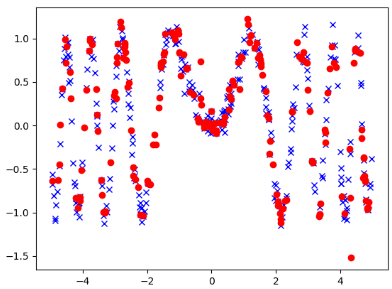

Figure 5.3: $\tau = 0.05$

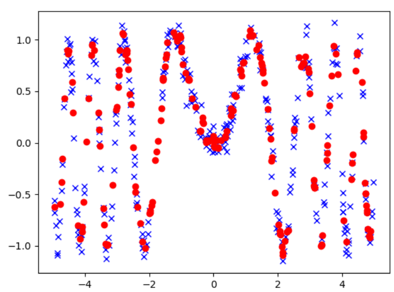

Figure 5.4: $\tau = 0.1$

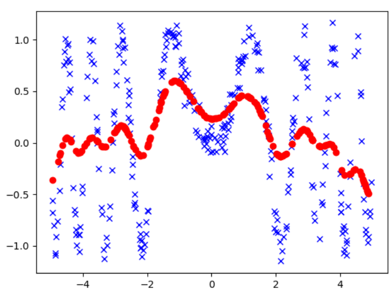

Figure 5.5: $\tau = 0.5$

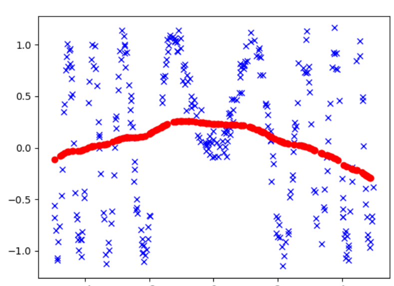

Figure 5.6: $\tau = 1.0$

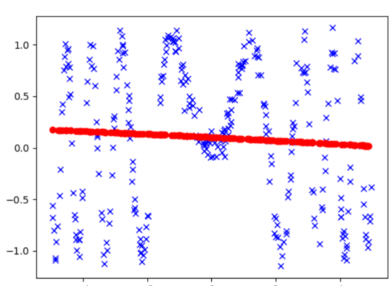

Figure 5.7: $\tau = 10.0$