Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

Assignment Number: 2
Student Name: Deepanshu Bansal
Roll Number: 150219
Date: October 10, 2017

QUESTION 1

1. No. It is not useful in learning a binary classifier from this data because it just defines different persons same as different S.No. and has nothing to do with good or bad advisor. Moreover each name is different and those cannot be classified into different classes. If we try to do so that would mean as many classes as many names. So not at all useful.

2. No. Consider S.No. entries "6" and "14" they have same values for all decision making attributes but still different final result. One is good advisor while other is not. So no matter how complicated the classification algorithm is we cannot perfectly classify this data.

3. The decision tree is as follows:

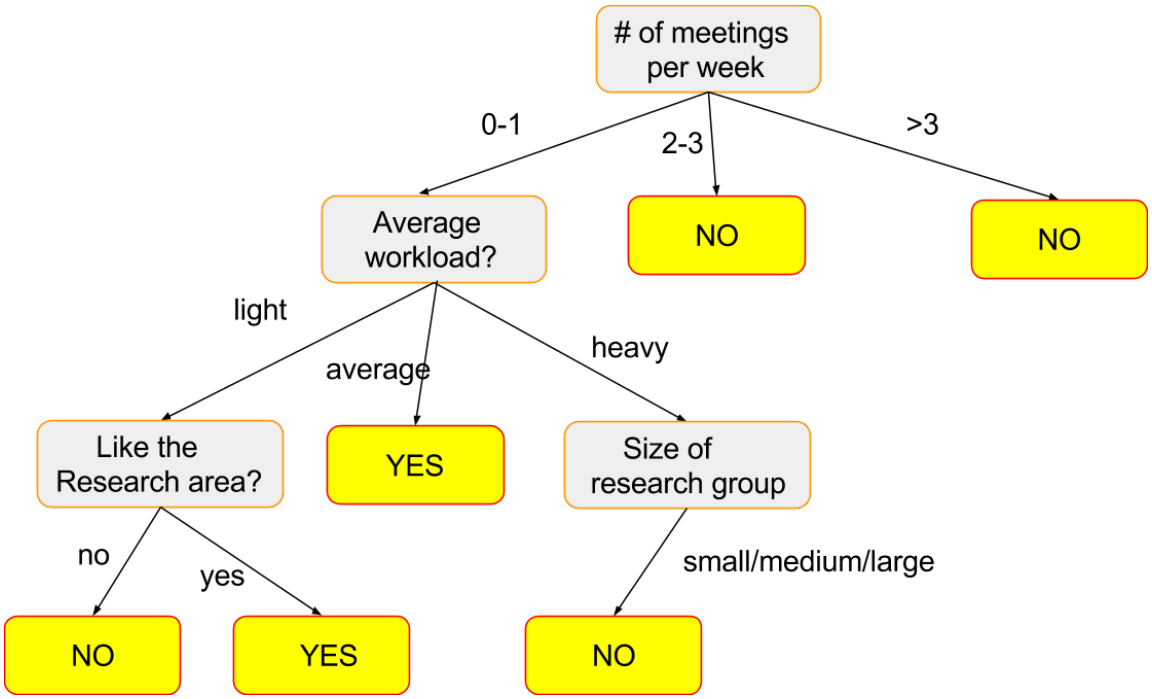

Figure 1: Decision Tree for given data.

Justification

1. I chose root node to be # of meetings per week because it has the highest information gain of 0.251.

   - $Gain(S$,Size of Research Group$) = 0.068$
   - $Gain(S$,Like Research Area$) = 0.032$
   - $Gain(S$,Workload$) = 0.061$

- $Gain(S$,no. of meetings$) = 0.251$

Since it has three possible values, the root node has three branches (0-1, 2-3, >3). Now the node corresponding to (0-1) branch is for Average Workload? because it has highest information gain of 0.439

- $Gain(S_{no.of meetings=0-1}$,Size of Research Group$) = 0.115$
- $Gain(S_{no.of meetings=0-1}$,Like Research Area$) = 0.035$
- $Gain(S_{no.of meetings=0-1}$,Workload$) = 0.439$

Now the node corresponding to (2-3) and (>3) branches are leaves beacause gain for them is zero.

- $Gain(S_{no.of meetings=2-3}$,Size of Research Group$) = 0.000$
- $Gain(S_{no.of meetings=2-3}$,Like Research Area$) = 0.000$
- $Gain(S_{no.of meetings=2-3}$,Workload$) = 0.000$
- $Gain(S_{no.of meetings>3}$,Size of Research Group$) = 0.000$
- $Gain(S_{no.of meetings>3}$,Like Research Area$) = 0.000$
- $Gain(S_{no.of meetings>3}$,Workload$) = 0.000$

For the splitting of Average Workload? node according to "light" information gain is highest for Like the research area? as 1.000 hence splitted to it.

- $Gain(S_{LightWorkload\_no.of meetings=0-1}$,Size of Research Group$) = 0.000$
- $Gain(S_{LightWorkload\_no.of meetings=0-1}$,Like Research Area$) = 1.000$

For "average" answer is sure to be "yes" as all three yes and for "heavy" it splitted to Size of research group as it has more information gain of 0.170

- $Gain(S_{HeavyWorkload\_no.of meetings=0-1}$,Size of Research Group$) = 0.170$
- $Gain(S_{HeavyWorkload\_no.of meetings=0-1}$,Like Research Area$) = 0.072$

2. I chose not to split nodes emerging from root with branch (2-3) and (>3) as leaves because **information gain** from them is **zero** and decision is sure to be NO. For the leaf emerging from "Average Workload?" it is so because at this point it have only 3 Yes hence making sure as YES leaf. Now for "Like research area" attribute only two points are given and only for "medium" value for "Size of research group" attribute we cant really decide but now it is decided on the basis of whether now the value of "Like research area?" is yes or no, as given in training if it's yes then result is yes else no hence they are two leaves. Similarly for "Size of research group" attribute answer is always no as it have 4 No's and 1 Yes hence resulting output as No and into a leaf.

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

Assignment Number: 2
Student Name: Deepanshu Bansal
Roll Number: 150219
Date: October 10, 2017

QUESTION 2

1. Here we use 2^L models as:

$$\mathbf{W}^0 = \left\{ \mathbf{w}^{\{0,0\}}, \mathbf{w}^{\{0,1\}}, ..., \mathbf{w}^{\{0,2^L-1\}} \right\}$$

Generative Story is

- Latent variable: z^i
- Observed variable: b^i
- Noise is given as gaussian with variance σ^2

Here z^i refers to which model means which subset of L is being used. We will try to minimize the noise using these models according to alternating algorithm. We can calculate bill as follows:

$$b^i = \left\langle \mathbf{w}^{z^j}, \mathbf{x}^i \right\rangle + \epsilon^i$$

$$\epsilon^i \sim \mathcal{N}(0, \sigma^2)$$

We will see for which particular model we can get minimum error for that corresponding bill.

2. Likelihood Expression: $\mathbb{P}\left[b \,|\, \mathbf{x}, \mathbf{W}\right]$

$$\mathbb{P}\left[b \,|\, \mathbf{x}, \mathbf{W}\right] = \prod_{i=1}^{n} \mathbb{P}\left[b^i \,|\, \mathbf{x}^i, \mathbf{W}\right]$$

$$\mathbb{P}\left[b^i \,|\, \mathbf{x}^i, \mathbf{W}\right] = \sum_{j=1}^{2^L} \mathbb{P}\left[b^j \,|\, \mathbf{x}^i, \mathbf{W}, z^j\right] . \mathbb{P}\left[z^j \,|\, \mathbf{x}^i, \mathbf{W}\right]$$

$$\mathbb{P}\left[b \,|\, \mathbf{x}, \mathbf{W}\right] = \prod_{i=1}^{n} \sum_{j=1}^{2^L} \mathbb{P}\left[b^j \,|\, \mathbf{x}^i, \mathbf{W}, z^j\right] . \mathbb{P}\left[z^j \,|\, \mathbf{x}^i, \mathbf{W}\right]$$

$$\mathbb{P}\left[z^j \,|\, \mathbf{x}^i, \mathbf{W}\right] = \frac{1}{2^L}$$

$$\mathbb{P}\left[b^j \,|\, \mathbf{x}^i, \mathbf{W}, z^j\right] = \mathcal{N}(\left\langle \mathbf{w}^{z^j}, \mathbf{x}^i \right\rangle, \sigma^2)$$

$$\mathbb{P}\left[b \,|\, \mathbf{x}, \mathbf{W}\right] = \frac{1}{2^L} \prod_{i=1}^{n} \sum_{j=1}^{2^L} \mathcal{N}(\left\langle \mathbf{w}^{z^j}, \mathbf{x}^i \right\rangle, \sigma^2)$$

$$\mathbb{P}\left[b \,|\, \mathbf{x}, \mathbf{W}\right] = \frac{1}{2^L} \prod_{i=1}^{n} \sum_{j=1}^{2^L} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(b^i - \left\langle \mathbf{w}^{z^j}, \mathbf{x}^i \right\rangle)}{2\sigma^2})^2$$

3. Alternating Algorithm

   (a) Initialize $\mathbf{W}^0 = \left\{\mathbf{w}^{\{0,0\}}, \mathbf{w}^{\{0,1\}}, ..., \mathbf{w}^{\{0,2^L-1\}}\right\}$.

   (b) For $i \in [n]$, update $z^{i,t}$ using $\mathbf{W}^t$

      i. Let $z^{i,t} = \underset{k \in 0,1,...,2^l}{\arg\min} |b^i - \langle \mathbf{w}^{t,k}, \mathbf{x}^i \rangle|$

   (c) Update $\mathbf{w}^{t+1,k} = \underset{\mathbf{w}}{\arg\min} \sum_{i:z^{i,t}=k} (b^i - \langle \mathbf{w}^{t,k}, \mathbf{x}^i \rangle)^2 + \frac{1}{2} \|w\|^2$

   (d) Set $\mathbf{W}^{t+1} = \left\{\mathbf{w}^{\{t+1,0\}}, \mathbf{w}^{\{t+1,1\}}, ..., \mathbf{w}^{\{t+1,2^L-1\}}\right\}$.

   (e) Repeat until convergence.

   Some points:

   - In step i of (b) we are assigning a label $z^{i,t}$ a value for which the bill closely belongs to that class ie. for which subset our error is minimum.

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

Assignment Number: 2
Student Name: Deepanshu Bansal
Roll Number: 150219
Date: October 10, 2017

QUESTION 3

We have the following problem:

$$\begin{aligned} \underset{\mathbf{w},\{\xi_i\}}{\arg\min}\ & \|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \xi_i^2 \\ \text{s.t. } & y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 1 - \xi_i, \text{ for all } i \in [n] \\ & \xi_i \geq 0, \text{ for all } i \in [n] \end{aligned} \qquad (P1)$$

1. We know that $\xi^2 \geq 0$ for all real ξ. Here we are trying to minimize the optimization problem hence we want all ξ_i to be equal to zero if constraints weren't there. Now we have constraint :

$$\begin{aligned} y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle &\geq 1 - \xi_i, \text{ for all } i \in [n] \\ \xi_i &\geq 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle, \text{ for all } i \in [n] \end{aligned}$$

Case-I: $y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 1$, we get

$$\begin{aligned} 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle &\geq 0, \text{ for all } i \in [n] \\ \xi_i &\geq 0, \text{ for all } i \in [n] \end{aligned}$$

Here we get constraints $\xi_i \geq 0$ for all $i \in [n]$ from the first constraint itself hence specifying constraints $\xi_i \geq 0$ for all $i \in [n]$ is vacuous in this case.

Case-II: $y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle > 1$, we get

$$\begin{aligned} 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle &< 0, \text{ for all } i \in [n] \\ \text{Let } 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle &= \alpha_i, \text{ for all } i \in [n] \\ \text{Now } \xi_i &\geq \alpha_i, \text{ for all } i \in [n] \end{aligned}$$

Since $\alpha_i < 0$ hence this means that ξ_i is greater than a negative value for all $i \in [n]$. Since we are minimizing the optimization problem we will always choose $\xi_i = 0$ for all $i \in [n]$ thus minimizing target as square of negative number is also greater then zero and we don't want that and thus the constraints $\xi_i \geq 0$ for all $i \in [n]$ are vacuous in this case also.

So the constraints $\xi_i \geq 0$ are vacuous i.e. the optimization problem does not change even if we remove the all constraints $\xi_i \geq 0$ for all $i \in [n]$.

2. Without vacuous constraints optimization problem becomes

$$\begin{aligned}\underset{\mathbf{w},\{\xi_i\}}{\arg\min}\ \ & \|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \xi_i^2 \\ \text{s.t. } & 1 - \xi_i - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0, \text{ for all } i \in [n]\end{aligned}$$

Lagrangian using multipliers α_i's s.t $\alpha_i \geq 0$ for all $i \in [n]$ is

$$L(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \xi_i^2 + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)$$

Now the optimization problem becomes

$$\underset{\mathbf{w},\{\xi_i\}}{\arg\min} \left\{ \underset{\boldsymbol{\alpha} \geq 0}{\arg\max} \{L(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha})\} \right\}$$

3. Derivation for the Lagrangian dual problem for $(P1)$ is as follows

$$\begin{aligned}\nabla_{\mathbf{w}} L &= 2\mathbf{w} - \sum_{i=1}^{n} \alpha_i y^i \mathbf{x}^i \\ \nabla_{\mathbf{w}} L &= 0 \\ \mathbf{w} &= \frac{1}{2} \sum_{i=1}^{n} \alpha_i y^i \mathbf{x}^i\end{aligned}$$

$$L(\boldsymbol{\xi}, \boldsymbol{\alpha}) = \left\| \frac{1}{2} \sum_{i=1}^{n} \alpha_i y^i \mathbf{x}^i \right\|_2^2 + \sum_{i=1}^{n} \xi_i^2 + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y^i \left\langle \frac{1}{2} \sum_{i=1}^{n} \alpha_i y^i \mathbf{x}^i, \mathbf{x}^i \right\rangle)$$

$$L(\boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{4} \sum_{n} \sum_{m} \alpha_n \alpha_m y^n y^m \left\langle \mathbf{x}^n, \mathbf{x}^m \right\rangle + \sum_{i=1}^{n} \xi_i^2 + \sum_{i=1}^{n} \alpha_i (1 - \xi_i) - \frac{1}{2} \sum_{n} \sum_{m} \alpha_n \alpha_m y^n y^m \left\langle \mathbf{x}^n, \mathbf{x}^m \right\rangle$$

$$L(\boldsymbol{\xi}, \boldsymbol{\alpha}) = -\frac{1}{4} \sum_{n} \sum_{m} \alpha_n \alpha_m y^n y^m \left\langle \mathbf{x}^n, \mathbf{x}^m \right\rangle + \sum_{i=1}^{n} (\xi_i^2 + \alpha_i - \alpha_i \xi_i)$$

$$\begin{aligned}\nabla_{\xi_i} L &= 2\xi_i - \alpha_i, \text{ for all } i \in [n] \\ \nabla_{\xi_i} L &= 0, \text{ for all } i \in [n] \\ \xi_i &= \frac{\alpha_i}{2}, \text{ for all } i \in [n]\end{aligned}$$

$$L(\boldsymbol{\alpha}) = -\frac{1}{4} \sum_{n} \sum_{m} \alpha_n \alpha_m y^n y^m \left\langle \mathbf{x}^n, \mathbf{x}^m \right\rangle + \sum_{i=1}^{n} (\alpha_i - \frac{\alpha_i^2}{4})$$

Thus Lagrangian dual problem for $(P1)$ is $\underset{\boldsymbol{\alpha} \geq 0}{\arg\max} \{L(\boldsymbol{\alpha})\}$ or $\underset{\boldsymbol{\alpha} \geq 0}{\arg\min} \{-L(\boldsymbol{\alpha})\}$

$$\underset{\boldsymbol{\alpha} \geq 0}{\arg\min} \left\{ \frac{1}{4} \sum_{n} \sum_{m} \alpha_n \alpha_m y^n y^m \left\langle \mathbf{x}^n, \mathbf{x}^m \right\rangle - \sum_{i=1}^{n} (\alpha_i - \frac{\alpha_i^2}{4}) \right\}$$

4. The original SVM problem is

$$\arg\min_{0\leq\boldsymbol{\alpha}\leq C}\left\{\frac{1}{2}\sum_{n}\sum_{m}\alpha_n\alpha_m y^n y^m\left\langle \mathbf{x}^n,\mathbf{x}^m\right\rangle-\sum_{i=1}^{n}\alpha_i\right\}$$

   The differences are

   (a) The constant in the term $\sum_n\sum_m \alpha_n\alpha_m y^n y^m \langle \mathbf{x}^n,\mathbf{x}^m\rangle$.
   We have $\frac{1}{4}$ while the original SVM problem has $\frac{1}{2}$.

   (b) There is extra $\sum_{i=1}^{n}\frac{\alpha_i^2}{4}$ term in dual problem for $(P1)ie.heredualproblemdependsonsquareofmultipl$

   (c) We dont get any other constraint on α. We have only $\boldsymbol{\alpha}\geq 0$ while original have $0\leq\boldsymbol{\alpha}\leq C$.

5. No, the positivity constraints $\xi_i\geq 0$ are not vacuous for the original SVM problem. If this constraint is removed then we may have a case where some points's labels are predicted correctly with high surety and while for others they are predicted wrong with low surety. Now even if number of first type of points is much smaller than the second type model will still think it is good because it will try to minimize $\sum_{i=1}^{n}\xi_i$ since $\xi_i\geq 1-y^i\langle\mathbf{w},x\rangle$ points with higher surety will lead to more negative values for ξ than with lower surety ones thus trying to minimize $\sum_{i=1}^{n}\xi$. Hence if $\xi_i\geq 0$ constraints is avoided it will lead to wrong results.

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

Assignment Number: 2
Student Name: Deepanshu Bansal
Roll Number: 150219
Date: October 10, 2017

QUESTION 4

1. Submitted in dropbox.

2. Submitted in dropbox.

3. In this case current iterate $\mathbf{w}^t$ give slightly better performance than the averaged iterate.

4. I selected step length by hit-and-trial method and it worked out well if step-length is inversely proportional to input size n. I took it as:

$$\eta = \frac{-10}{n\sqrt{t+1}}$$

5. Required graphs are as shown. The first graph is for Vanilla gradient descent, middle is for stochastic coordinate descent and last one is combination of both. These graphs are plotted for 1000 points ie. for vanilla gradient descent it took 10,000 iterations with 10 spacing and for stochastic coordinate descent it took 100,000 iterations with 100 spacing.

   It is clear from the graphs that SCD reduce the objective value $f(\mathbf{w})$ faster than GD.

6. There isn't much difference between graphs based on two types of time spent. Graph for theoretical time have much lager values of x as compared to time_elapsed. Moreover graphs for SCD reaches much faster to optimum than for GD. Even though GD is slower it gives slightly good value than SCD as can be seen from the graph. We can see that graph for GD is very slow hence very little (shown in purple) as compared to SCD.

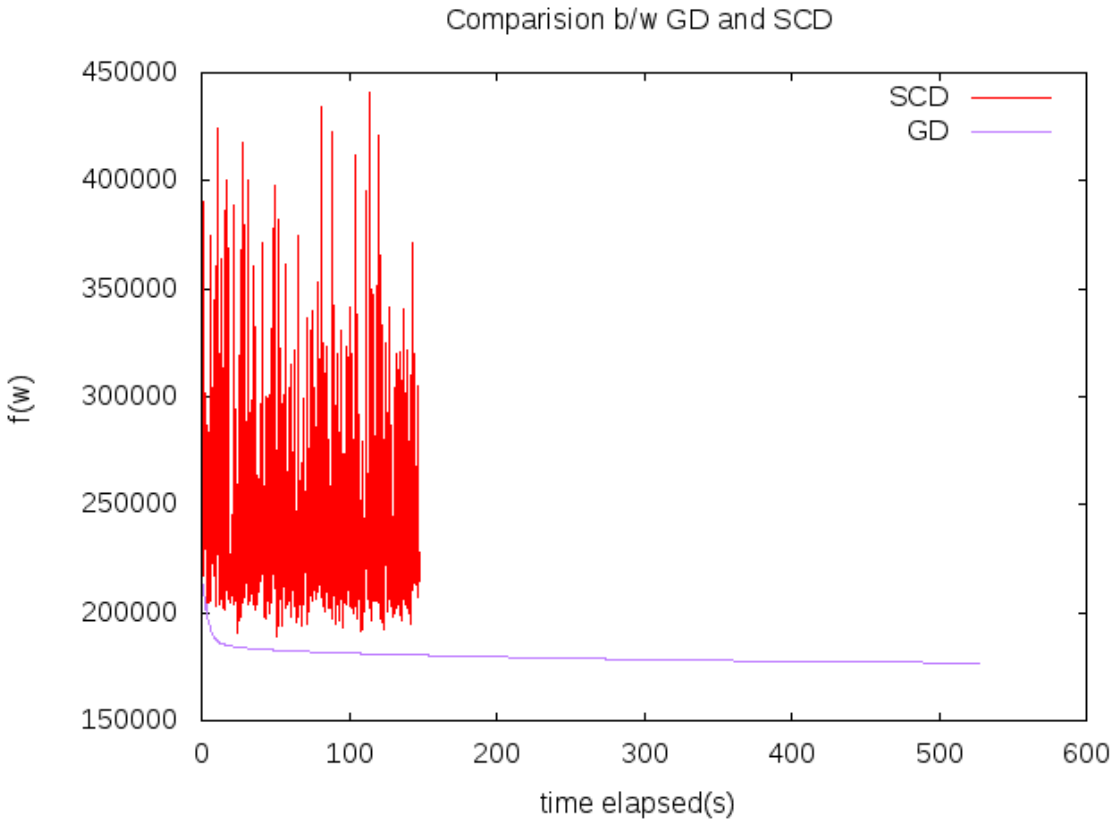

Figure 2: Gradient Descents (GD) and (SCD)

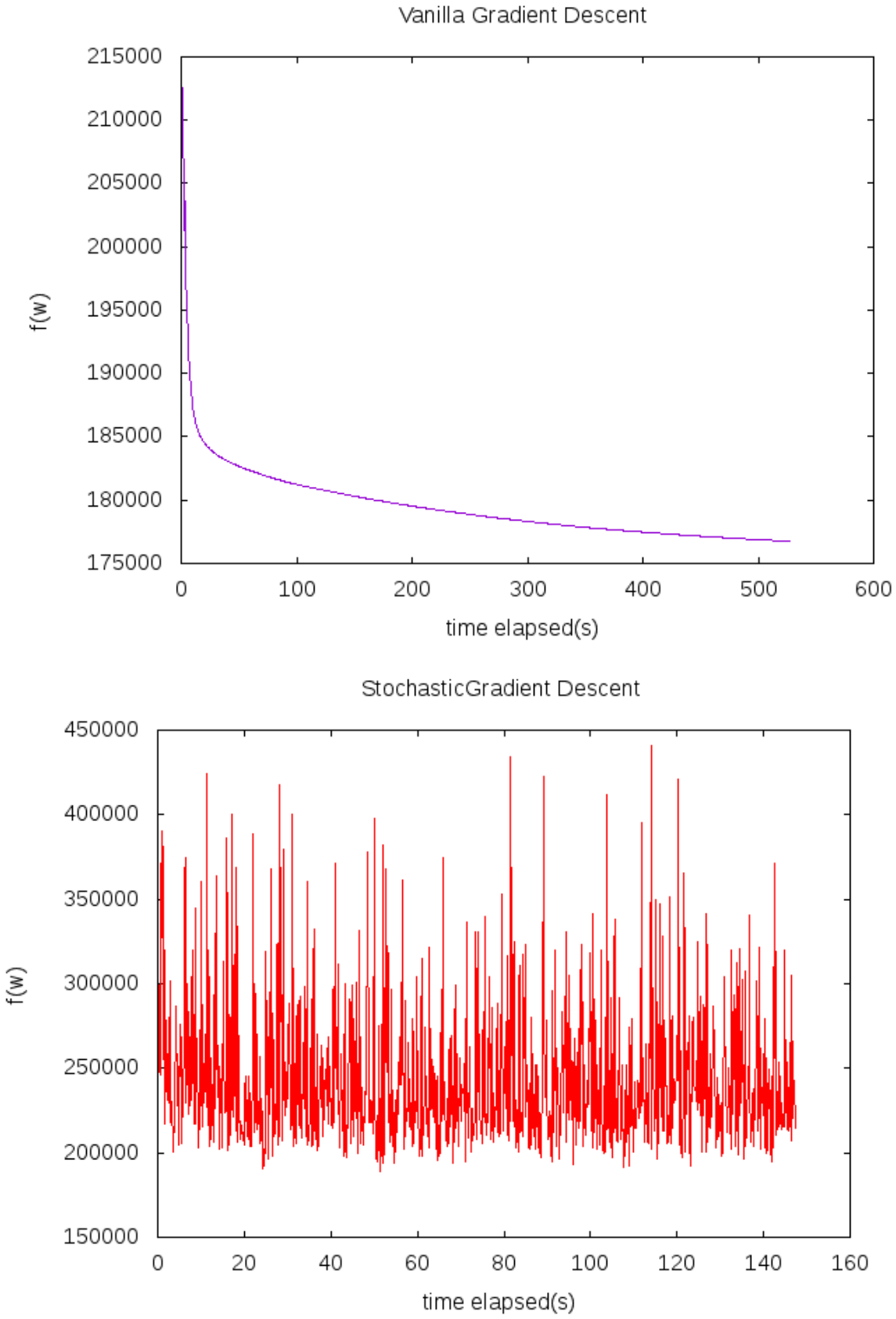

Figure 3: GD and SCD on different graphs

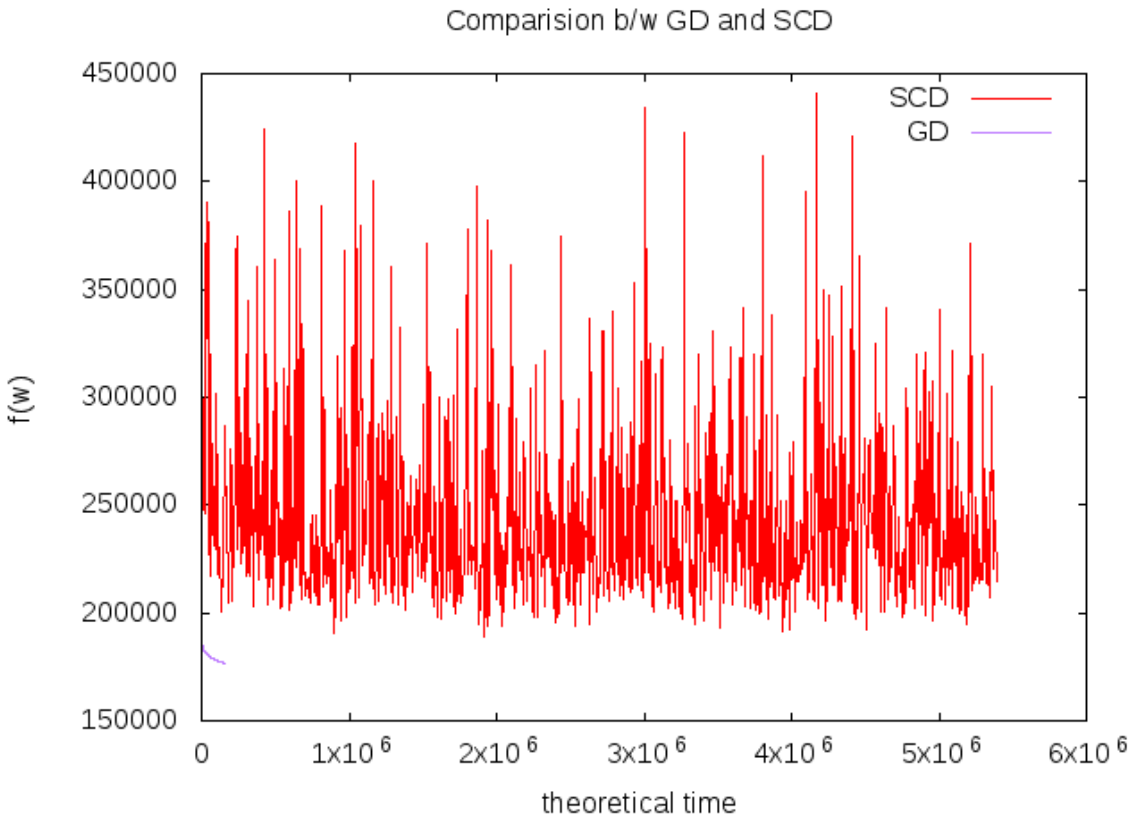

Figure 4: GD and SCD vs theoretical time on same graph

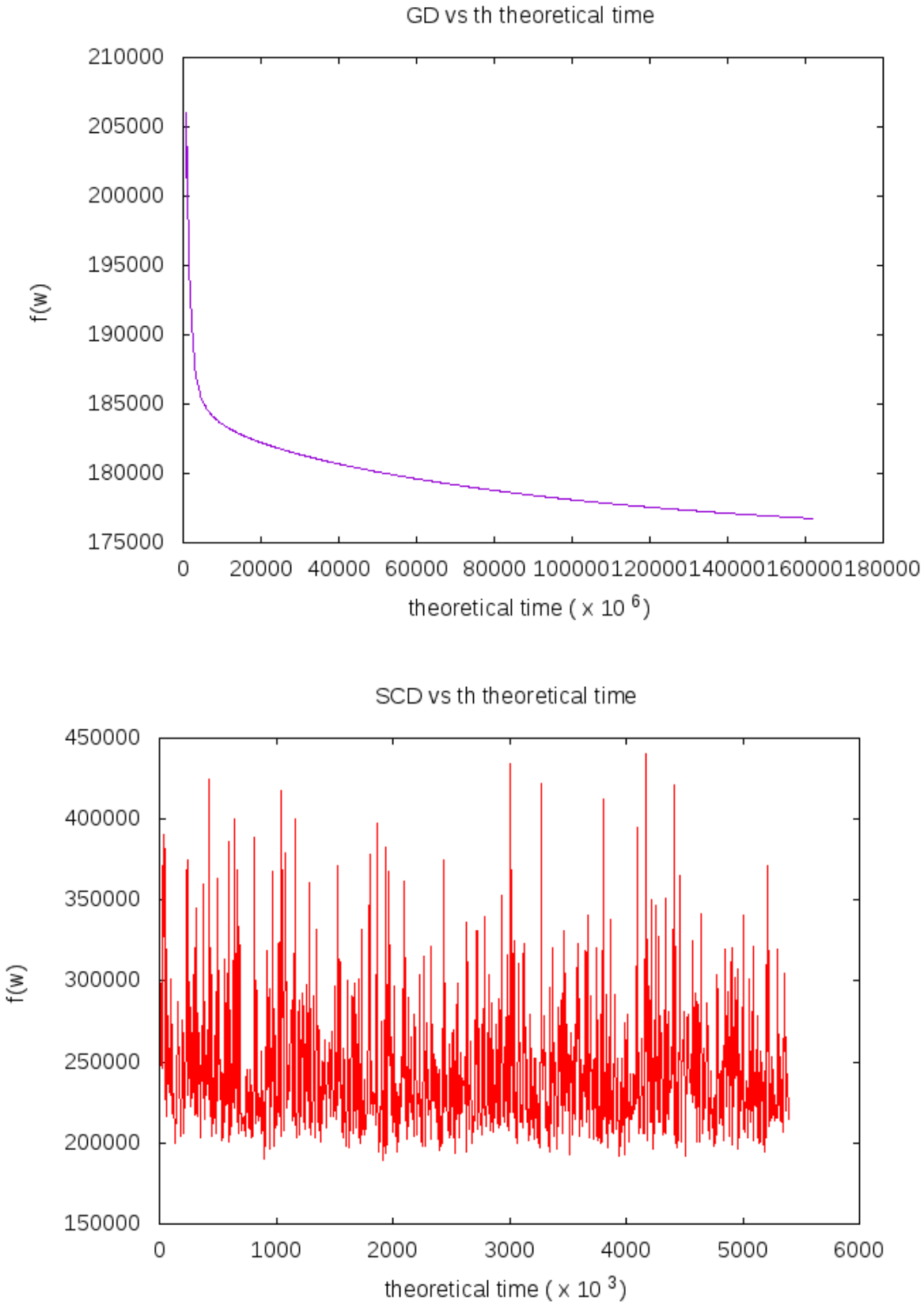

Figure 5: GD and SCD vs theoretical time on different graphs