Machine Learning, Spring 2018
Homework 1

Due on 23:59 Mar 15, 2018
Send [redacted]
with subject "Chinese name+student number+HW1"

1 Preliminaries

1) Give at least 2 examples of machine learning applications in your life. For each example, please describe how you think this real-world application can be reduced to a machine learning problem.

2) f is twice continuously differentiable. At a point $x \in \mathbb{R}^n$, direction d is a descent direction, i.e., $\nabla f(x)^T d < 0$. Show that we can decrease f by moving (a sufficiently small distance) along such a direction.

2 Understanding Convex function and First-order necessary condition

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function defined in a convex set and that $p \in \Re^n$,

1) Show that f is convex if the Hessian of f is positive semidefinite.
(Key: $f(x) = f(x') + \nabla f(x')^T(x - x') + \frac{1}{2}(x - x')^T H(x' + \lambda(x - x'))(x - x')$, one can conclude that $f(x) \geq f(x') + \nabla f(x')^T(x - x')$ using the property of Hessian.)

2) Suppose f is convex, show that the global minimizer satisfies the condition $\nabla f(x^*) = 0$.
(Key: Using the fact $\nabla f(x^*)^T(x-x^*) \geq 0$ and $f(x) \geq f(x^*)+\nabla f(x^*)^T(x-x^*)$.)

3 Linear Regression via Gradient Descent Method

How strong is the linear relationship between the age of a driver and the distance the driver can see? If we had to guess, we might think that the relationship is negative as age increases, the distance decreases. A research firm collected data on a sample of n = 30 drivers, where the data is provided in "**data.txt**".

1) Formulate this problem with the linear regression and give its expression. Give the expression of the cost function $J(\boldsymbol{\theta})$.

2) Use the gradient descent method to solve this linear regression problem, and design a termination criterion. Plot your result.

Notice:

Please finish your simulation with MATLAB/Python and compress your codes into one file and sent it to TAs. (Data preprocessing is recommended, otherwise the algorithm may crash, i.e., Age/10, Distance/100.)

In pseudocode, the gradient descent method can be presented as follows:

Algorithm 1 Gradient descent

```
Given the desired accuracy ϵ.
Initialize the parameter θ and the learning rate α.
repeat
    Update θ := θ − α∇J(θ).
until Your termination criterion.
return θ.
```

4 How to Deal with Outliers

In the "Lecture 3: Linear Regression, Gradient Descent", we sum up the squares of the differences between the actual value and the estimated value

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h(x^i) - y^i)^2.$$

This error function is the one most frequently used, but it is one of several possible error functions. Because it sums up the squares of the differences, it is not robust to outliers. we just add two outliers in a data for linear regression, assume that for the large data set, you could not find the outliers manually, what would be a better error function to implement robust regression?

We need you to show that your error function could reduce the loss of test data compared with the square loss. The data is provided in "**5-trainingdata.txt**" and "**5-testdata.txt**".