HW Assignment 2, Theory

Problems marked with a $(*)$ are mandatory only for ITCS 8156 students. Bonus problems are optional, solving them will result in extra points.

1 Basic Calculus (10 + 5 points)

By setting the gradient to 0, find the solutions to the following optimization problems:

1. (5 points) $\hat{x} = \arg\min_{x} J(x)$,where $J(x) = x^2 - 2x + 3$.
2. (5 points) $\hat{x}, \hat{y} = \arg\min_{x,y} J(x, y)$, where $J(x, y) = 2x^2 + 3y^2 - 4x + 12y + 15$.
3. (5 points) $(*)$ $\hat{x}, \hat{y} = \arg\min_{x,y} J(x, y)$, where $J(x, y) = x^2 + 4y^2 - 4xy + 2x - 4y + 4$.

1.1 Convexity (5 bonus points)

Prove that $J(x, y)$ at point 2 above is convex.

2 Basic Linear Algebra (20 + 10 points)

Using the definition of the dot-product, transposition, and multiplication operators on matrices, prove the simple identities below, where $\mathbf{a}, \mathbf{b}$ are two column vectors $n \times 1$, A and B are two $m \times n$ matrices, and $c \in R$ is a constant:

1. (5 points) $\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\mathbf{a}$.
2. (5 points) $(A + B)^T = A^T + B^T$.
3. (5 points) $(A^T B)^T = B^T A$.
4. (5 points) $(cA)^T = cA^T$.
5. (10 points) $(*)$ Prove that the Moore-Penrose inverse of a square matrix A is equal with the inverse matrix A^{-1}.

3 Polynomial Curve Fitting (20 points)

Consider the problem of fitting a dataset of N points with a polynomial of degree M, by minimizing the sum-of-squares error:

$$J(\mathbf{w}) = \frac{1}{2N} \sum_{n=1}^{N} (h_{\mathbf{w}}(\mathbf{x}_n) - t_n)^2 \tag{1}$$

where $h_{\mathbf{w}}(\mathbf{x}) = \sum_{j=0}^{M} w_j x^j$.

By setting the gradient to zero, show that the solution to minimizing $J(\mathbf{w})$ satisfies the following set of linear equations:

$$\sum_{j=0}^{M} A_{ij} w_j = T_i \tag{2}$$

$$\text{where } A_{ij} = \sum_{n=1}^{N} x_n^{i+j} \text{ and } T_i \ = \ \sum_{n=1}^{N} x_n^i t_n \tag{3}$$

3.1 Adding L_2 regularization (∗) (10 points)

Derive the solution for the regularized version of polynomial curve fitting, which minimizes the objective function below:

$$J(\mathbf{w}) = \frac{1}{2N} \sum_{n=1}^{N} (h_{\mathbf{w}}(\mathbf{x}_n) - t_n)^2 + \frac{\lambda}{2} ||\mathbf{w}||^2 \tag{4}$$

Make sure that your derivation of the solution ends with the vectorized version shown in class, i.e. using matrices and vectors.

3.2 Adding L_2 regularization (10 + 10 bonus points)

Consider again the regularized linear regression objective:

$$J(\mathbf{w}) = \frac{1}{2N} \sum_{n=1}^{N} (h(\mathbf{x}_n, \mathbf{w}) - t_n)^2 + \frac{\lambda}{2} ||\mathbf{w}||^2$$

1. (10 points) Minimizing the L2 norm of $\mathbf{w}$ drives all parameters, including w_0, towards 0. Are there situations in which we do not want to constrain w_0 to be small? If yes, give an example, if not show why you think it is useful to constrain all the weights to be small, including w_0 .

2. (10 bonus points) We have seen in class how to compute the weights $\mathbf{w}$ that minimize $J(\mathbf{w})$. Assume now that we replace $||\mathbf{w}||$ in $J(\mathbf{w})$ with $||\mathbf{w}_{[\mathbf{1}:]}||$, where $\mathbf{w}_{[\mathbf{1}:]} = [w_1, w_2, ..., w_M]$. Derive the solution for $\mathbf{w}$ that minimizes this new objective function.

4 Bonus (10 points)

Consider the regression problem discussed on slide 4 in the introduction lecture, where input data were noisy measurements of speed and time, and the labels were noisy measurements of distance. If a multiple linear regression were trained on such data, would it be able to make good predictions? If yes, explain why. If not, explain how you would transform the input data so that a multiple linear regression model trained on it made good predictions.

5 Submission

Submit your responses on Canvas as one file named `theory.pdf`. It is recommended to use an editor such as Latex or Word or Jupyter-Notebook that allows editing and proper formatting of equations. Alternatively, if you choose to write your solutions on paper, submit an electronic scan / photo of it on Canvas. Make sure that your writing is legible and easy to read.