Machine Learning, Tutorial 2
University of Bern

Mehdi Noroozi (noroozi@inf.unibe.ch)

26/09/2018

1 Optimization and Least Mean Squares

1. Consider the least mean square problem:

$$\min_{x \in \mathbb{R}^n} \quad ||Ax - b||_2^2$$

(a) Suppose $A \in \mathbb{R}^{m \times n}$ is a full rank matrix and $m \geq n$. Find the closed-form solution of the east mean square problem.
Hint. If $A \in \mathbb{R}^{m \times n}$ is a full rank matrix and $m \geq n$, then $A^\top A$ is a positive definite matrix.
Solution.

$$||Ax - b||_2^2 = (Ax - b)^\top (Ax - b) = x^\top A^\top Ax - 2b^\top Ax + b^\top b$$

$$\begin{aligned} \nabla_x (x^\top A^\top Ax - 2b^\top Ax + b^\top b) &= \nabla_x x^\top A^\top Ax - \nabla_x 2b^\top Ax + \nabla_x b^\top b \\ &= 2A^\top Ax - 2A^\top b \end{aligned}$$

We know that $A^\top A$ is positive definite and invertible. Setting this last expression equal to zero and solving for x we have $x = (A^\top A)^{-1} A^\top b$.

(b) Suppose that A is not full rank. Write down the gradient descent step for the optimization problem. Is it guaranteed for gradient descent to converge to the global optimum?
Solution.

$$x_{t+1} := x_t - 2\alpha(A^\top Ax - A^\top b)$$

Suppose a function $f : \mathcal{R}^n : \mathcal{R}$ is twice differentiable.Then f is convex if and only if its Hessian is positive semidefinite, i.e. $\nabla_x^2 f(x) \geq 0$. We have $\nabla_x^2 ||Ax - b||_2^2 = 2A^\top A$ and we know that $A^\top A$ is a positive semidefinite matrix. Therefore, the least square objective function is convex. For a convex optimization problem all locally optimal points are globally optimal. Therefore, gradient descent converges to the global optimum of the least mean square problem.

2. Show that the solution of the following equally constrained problem is an eigenvector of A. Where $A \in \mathbb{S}^n$ is a symmetric matrix.

$$\max_{x \in \mathbb{R}^n} x^\top Ax \quad \text{subject to} \quad ||x||_2^2 = 1$$

Solution. A standard way of solving optimization problems with equality constraints is by forming the **Lagrangian**, and objective function that includes the equality constraints. The Lagrangian in this case can be given by:

$$\mathcal{L}(x, \lambda) = x^\top Ax - \lambda x^\top x$$

Where λ is called the Lagrangian multiplierassociated with the equality constraines. It can be established that for x^* to be an optimal to the problem, the gradient of the Lagrangian has to be zero at x^*. That is,

$$\nabla_x(\mathcal{L}(x,\lambda)) = \nabla_x(x^\top Ax - \lambda x^\top x) = 2Ax - 2\lambda x = 0.$$

This shows that the only point which can be possibly maximize (or minimize) $x^\top Ax$ assuming $x^\top x = 1$ are the eigenvectors of A.

3. We assumed that the hypothesis function has the form

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x.$$

Consider the case when the hypothesis is instead in the form

$$h_\theta(x) = \sum_{i=0}^{m} \theta_i \phi_i(x) = \theta^T \phi(x),$$

where ϕ is an arbitrary feature map. Consider the LMS prediction error:

$$J(\theta) = \frac{1}{2}\sum_{k=0}^{N} (\theta^\top \phi(x_j) - y_j)^2$$

- Work out the gradient descent step for this new hypothesis function.
Solution. *For one training sample the error is $J(\theta) = \frac{1}{2}(h_\theta(x) - y)^2 = \frac{1}{2}(\theta^T\phi(x) - y)^2$. The gradient descent step is $\theta_j := \theta_j - \alpha\frac{\partial J(\theta)}{\partial \theta_j}$.*

$$\begin{aligned} \tfrac{\partial J(\theta)}{\partial \theta_j} &= (h_\theta(x) - y)\tfrac{\partial}{\partial \theta_j}(h_\theta(x) - y) \\ &= (h_\theta(x) - y)\tfrac{\partial}{\partial \theta_j}(\sum_{i=0}^{m} \theta_i\phi_i(x) - y) \\ &= (h_\theta(x) - y)\phi_j(x) \end{aligned}$$

$$\theta_{j+1} := \theta_j + \alpha(y - h_\theta(x))\phi_j(x)$$

Notice, that the gradient descent step is very similar to the original case. We only needed to change x_j to $\phi_j(x)$.

- Is there an analytical solution for the LMS prediction in this case? If yes, compute the formula of the solution.
Solution. *Let F be a matrix, where each row contains $\phi(x^{(i)})^\top$, the transpose of feature representation of sample $x^{(i)}$. The error can be written in a form*

$$J(\theta) = \frac{1}{2}(F\theta - Y)^\top(F\theta - Y),$$

where Y is a column vector of the labels $y^{(i)}$. The parameters that minimise the error can be obtained by the following formula.

$$\theta = (F^\top F)^{-1}F^\top Y$$

Notice, that this formula is very similar to the original one, we only needed to change X to F.