Machine Learning, Tutorial 2
Universität Bern

Simon Jenni (simon.jenni@inf.unibe.ch)

Optimization and Least Mean Squares

1. Consider the least mean squares 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 least mean squares 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:
   Let us first expand the the objective function:

$$\begin{aligned} ||Ax - b||_2^2 &= (Ax - b)^\top (Ax - b) \\ &= x^\top A^\top Ax - x^\top A^\top b - b^\top Ax + b^\top b \\ &= x^\top A^\top Ax - 2x^\top A^\top b + b^\top b \end{aligned}$$

   This is a convex function of x and so to find the minimum we take the derivative and set it equal to zero:

$$\nabla_x (x^\top A^\top Ax - 2x^\top A^\top b + b^\top b) = 2A^\top Ax - 2A^\top b \stackrel{!}{=} 0$$

   We know that $A^\top A$ is positive definite and invertible. Solving the last equation 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:
   The gradient descent update step is given by

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

A twice differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if and only if the Hessian of f 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. This shows that the least squares 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 (provided a good choice of the learning rate α).

2. Suppose $A \in \mathbb{R}^{n \times n}$ is a symmetric matrix. Show that the solution of the following equality constrained optimization problem is an eigenvector of A.

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

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

$$\mathcal{L}(x, \lambda) = x^\top A x - \lambda(x^\top x - 1).$$

The parameter λ is called the Lagrangian multiplier associated with the equality constraint. It can be shown that for x^* to be an optimal solution to the problem, the gradient of the Lagrangian w.r.t. x has to be zero at x^*. That is,

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

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

3. The Linear Regression objective is given by

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{N} \left(h(x^{(i)}) - y^{(i)} \right)^2,$$

and we assumed that the hypothesis function has the form

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

Consider the case when the hypothesis is instead given by

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

where $\phi : \mathbb{R}^n \mapsto \mathbb{R}^m$ is an arbitrary feature map.

- Work out the gradient descent step for this new hypothesis function.

  Solution:

  For one training sample the error is given by $J(\theta) = \frac{1}{2}(h_\phi(x) - y)^2 = \frac{1}{2}(\theta^T\phi(x) - y)^2$. The gradient descent step is $\theta_j := \theta_j - \alpha\frac{\partial J(\theta)}{\partial \theta_j}$.

$$\begin{aligned}\frac{\partial J(\theta)}{\partial \theta_j} &= (h_\phi(x) - y)\,\frac{\partial}{\partial \theta_j}\,(h_\phi(x) - y)\\ &= (h_\phi(x) - y)\,\frac{\partial}{\partial \theta_j}\left(\sum_{i=0}^{m}\theta_i\phi(x)_i - y\right)\\ &= (h_\phi(x) - y)\,\phi(x)_j\end{aligned}$$

$$\theta_j := \theta_j + \alpha\,(y - h_\phi(x))\,\phi(x)_j$$

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

- 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.