EECS 545 – Machine Learning - Homework #1

David Ke Hong

Due: 11:00pm 01/25/2016

Homework Policy: Working in groups is fine, but each member must submit their own writeup. Please write the members of your group on your solutions. There is no strict limit to the size of the group but we may find it a bit suspicious if there are more than 4 to a team. Questions labelled with **(Challenge)** are not strictly required, but you'll get some participation credit if you have something interesting to add, even if it's only a partial answer. For coding problems, please report your results (values, plots, etc.) in your written solution. You will lose points if you only include them in your code submissions. Homework will be submitted via Gradescope (https://gradescope.com/).

1) **Linear Algebra (25 pts).**

(a) Are the following statements true or false? If true, prove it; if false, show a counterexample.

(i) Given an invertible matrix A, $(A^{-1})^\top = (A^\top)^{-1}$.

(ii) Given that matrix A, B, $A+B$ are invertible, $(A+B)^{-1} = A^{-1} + B^{-1}$.

(iii) The inverse of a symmetric matrix is itself symmetric.

(b) Singular value decomposition (SVD) factorizes a $m \times n$ matrix X as $X = U\Sigma V^\top$, where $U \in \mathbb{R}^{m\times m}$ and $U^\top U = UU^\top = I$, $\Sigma \in \mathbb{R}^{m\times n}$ contains non-increasing non-negative values along its diagonal and zeros elsewhere, and $V \in \mathbb{R}^{n\times n}$ and $V^\top V = VV^\top = I$. Given the SVD of a matrix $X = U\Sigma V^\top$, what is the eigendecomposition of $XX^\top$? (You need to define an appropriate square matrix Q and diagonal matrix Λ such that $XX^\top = Q\Lambda Q^{-1}$.)

(c) Matrix A is stored in `A.csv`. See `hw1.py` for details.

(i) Perform SVD on A and report the top 3 singular values. (No need to submit code this time.)

(ii) What happens if we zero out all but the top 3 singular values? Specifically, compute $\|A - B\|_F^2$ (the element-wise squared error between A and B), where B is defined as follows: given the SVD of $A = U\Sigma V^\top$, $B = U\Sigma_3 V^\top$, where Σ_3 is the same as Σ except all but the top 3 singular values are zero. (No need to submit code this time.)

2) **Probability (20 pts).**

(a) For the following equations, describe the relationship between them. Write one of four answers: "=", "≤", "≥", or "depends" to replace the "?". Choose the most specific relation that always holds and briefly explain why. Assume all probabilities are non-zero.

(i) $P(H = h|D = d)$? $P(H = h)$

(ii) $P(H = h|D = d)$? $P(D = d|H = h)P(H = h)$

(b) Random variables X and Y have a joint distribution $p(x, y)$. Prove the following results. You can assume continuous distributions for simplicity.

(i) $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]]$

(ii) $\mathrm{var[X]} = \mathbb{E}_Y[\mathrm{var}_X[X|Y]] + \mathrm{var}_Y[\mathbb{E}_X[X|Y]]$

3) **Positive (Semi-)Definite Matrices (20 pts).** Let A be a real, symmetric $d \times d$ matrix. We say A is *positive semi-definite* (PSD) if, for all $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x}^\top A\mathbf{x} \geq 0$. We say A is *positive definite* (PD) if, for all $\mathbf{x} \neq 0$, $\mathbf{x}^\top A\mathbf{x} > 0$. We write $A \succeq 0$ when A is PSD, and $A \succ 0$ when A is PD.

The *spectral theorem* says that every real symmetric matrix A can be expressed $A = U\Lambda U^\top$, where U is a $d\times d$ matrix such that $UU^\top = U^\top U = I$ (called an orthogonal matrix), and $\Lambda = diag(\lambda_1, ..., \lambda_d)$. Multiplying on the right by U we see that $AU = U\Lambda$. If we let $\mathbf{u}_i$ denote the i^{th} column of U, we have $A\mathbf{u}_i = \lambda_i\mathbf{u}_i$ for each i. This expression reveals that the λ_i are eigenvalues of A, and the corresponding columns $\mathbf{u}_i$ are eigenvectors associated to λ_i.

Using the spectral decomposition, show that

(a) A is PSD iff $\lambda_i \geq 0$ for each i.

(b) A is PD iff $\lambda_i > 0$ for each i.

4) **Maximum Likelihood Estimation (15 pts).** Consider a random variable $\mathbf{X}$ (possibly a vector) whose distribution (density function or mass function) belongs to a parametric family. The density or mass function may be written $f(\mathbf{x}; \theta)$, where θ is called the parameter, and can be either a scalar or vector. For example, in the Gaussian family, θ can be a two-dimensional vector consisting of the mean and variance. Suppose the parametric family is known, but the value of the parameter is unknown. It is often of interest to estimate this parameter from observations of $\mathbf{X}$.

Maximum likelihood estimation is one of the most important parameter estimation techniques. Let $\mathbf{X_1}, ..., \mathbf{X_n}$ be iid (independent and identically distributed) random variables distributed according to $f(\mathbf{x}; \theta)$. By independence, the joint distribution of the observations is the product

$$\prod_{i=1}^{n} f(\mathbf{X_i}; \theta) \tag{1}$$

Viewed as a function of θ, this quantity is called the likelihood of θ. It is often more convenient to work with the *log-likelihood*,

$$\sum_{i=1}^{n} \log f(\mathbf{X_i}; \theta) \tag{2}$$

A maximum likelihood estimate (MLE) of θ is any parameter

$$\hat{\theta} \in \arg\max_{\theta} \sum_{i=1}^{n} \log f(\mathbf{X_i}; \theta) \tag{3}$$

where "arg max" denotes the set of all values achieving the maximum. If there is a unique maximizer, it is called the maximum likelihood estimate. Let $\mathbf{X_1}, ..., \mathbf{X_n}$ be iid Poisson random variables with intensity parameter λ. Determine the maximum likelihood estimator of λ.

5) **Unconstrained Optimization (20 pts).** In this problem you will prove some of properties of unconstrained optimiziation problems.

(a) Show that if f is strictly convex, then f has at most one global minimizer.

For the next two parts, the following fact will be helpul. A twice continuously differentiable function admits the quadratic expansion

$$f(\mathbf{x}) = f(\mathbf{y}) + \langle \bigtriangledown f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2} \langle \mathbf{x} - \mathbf{y}, \bigtriangledown^2 f(\mathbf{y})(\mathbf{x} - \mathbf{y}) \rangle + o(\|\mathbf{x} - \mathbf{y}\|^2) \tag{4}$$

where $o(t)$ denotes a function satisfying $\lim_{t \to 0} \frac{o(t)}{t} = 0$, as well as the expansion

$$f(\mathbf{x}) = f(\mathbf{y}) + \langle \bigtriangledown f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{1}{2} \langle \mathbf{x} - \mathbf{y}, \bigtriangledown^2 f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))(\mathbf{x} - \mathbf{y}) \rangle \tag{5}$$

for some $t \in (0, 1)$.

(b) Show that if f is twice continuously differentiable and $\mathbf{x}^*$ is a local minimizer, then $\bigtriangledown^2 f(\mathbf{x}^*) \succeq 0$, *i.e.*, the Hessian of f is positive semi-definite at the local minimizer $\mathbf{x}^*$.

(c) Show that if f is twice continuously differentiable, then f is convex if and only if the Hessian $\bigtriangledown^2 f(\mathbf{x})$ is positive semi-definite for all $\mathbf{x} \in \mathbb{R}^d$.

(d) Consider the function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A \mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$, where A is a symmetric $d \times d$ matrix. Derive the Hessian of f. Under what conditions on A is f convex? Strictly convex?