Machine Learning for Finance – Problem Set 6 Solutions

Neal Parikh

May 1, 2018

Instructions. Do not refer to any outside sources to complete this assignment, in accordance with the honor code. If you discussed any problems with other students, indicate that in your solutions.

1. *Backtest timing.* Suppose the $T \times n$ asset returns matrix R gives the returns of n assets over T periods. When the n-vector w gives a set of portfolio weights, the T-vector Rw gives the time series of portfolio return over the T time periods. Evaluating portfolio return with past returns data is called *backtesting*.

   Consider a specific case with $n = 5000$ assets, and $T = 2500$ returns. (This is 10 years of daily returns, since there are around 250 trading days in each year.) About how long would it take to carry out this backtest on a 50 Gflop/s computer?

   Solution. A matrix-vector product Rw of an $T \times n$ matrix with an n-vector w takes $2Tn$ flops. For $n = 5000$ and $T = 2500$ this is 2.5×10^7 flops. On a 50 Gflop/s computer it takes about 500 microseconds.

2. *Gradient descent for logistic regression.* Derive the gradient descent update rule for solving the (unregularized) maximum likelihood estimation problem for a logistic regression model.

   Solution. The only work is to derive the gradient of the log likelihood, which is given by

   $$\nabla \ell(w) = \sum_{i=1}^{N} (y_i - s(w^T x_i)) x_i,$$

   where s is the sigmoid function. Then, for example, a stochastic gradient method update rule would be

   $$w^{k+1} := w^k + \alpha(y_i - s(w^T x_i)) x_i,$$

   where α is a step size and the training example (x_i, y_i) is chosen from the full training set $\mathcal{D}$ via some policy (uniformly at random, cyclically, *etc.*). (This is for a gradient *ascent* rule in which we are maximizing ℓ, as opposed to a descent rule for minimizing $-\ell$.)

3. *Linear invariance of optimization algorithms.* Suppose you would like to use a linear regression model to predict the price of houses. In your model, you use the features

   $$\begin{aligned} x_0 &= 1 \\ x_1 &= \text{size in square meters} \\ x_2 &= \text{height of roof in meters.} \end{aligned}$$

Suppose a colleague repeats the same analysis using the same training set, but uses the feature

$$x_2' = \text{height of roof in cm}$$

instead of x_2, *i.e.*, $x_2' = 100x_2$.

(a) Suppose you both fit a linear regression model via the normal equations. (Assume there are no degeneracies, so you obtain a unique solution.) You get parameters θ and he gets θ'. Is the following true?

$$\theta_0 = \theta_0', \quad \theta_1 = \theta_1', \quad \theta_2 = 100\theta_2'.$$

Solution. True. This follows from the invariance of Newton's method to linear reparameterizations, since running a single step of Newton's method for a linear regression problem is equivalent to solving the normal equations.

(b) Suppose you both run linear regression, initialize the parameters to zero, and compare results after running a single iteration of (batch) gradient descent. Does the relation between θ and θ' above hold?

Solution. False; gradient descent is not invariant to linear reparameterizations.

4. *k-means with nonnegative, proportions, or Boolean vectors.* Suppose that the vectors $x_1, \ldots, x_N$ are clustered using k-means, with group representatives $z_1, \ldots, z_k$.

(a) Suppose the original vectors x_i are nonnegative, *i.e.*, their entries are nonnegative. Explain why the representatives z_j are also nonnegative.

(b) Suppose the original vectors x_i represent proportions, *i.e.*, their entries are nonnegative and sum to one. (This is the case when x_i are word count histograms, for example.) Explain why the representatives z_j also represent proportions, *i.e.*, their entries are nonnegative and sum to one.

(c) Suppose the original vectors x_i are Boolean, *i.e.*, their entries are either 0 or 1. Give an interpretation of $(z_j)_i$, the ith entry of the j group representative.

Solution. The group representative z_j is the average of the vectors x_k in the group:

$$z_j = \frac{1}{|G_j|} \sum_{k \in G_j} x_k.$$

(a) If the vectors x_k are nonnegative, the average z_j is a nonnegative vector.

(b) If each vector sums to one, $\mathbf{1}^T x_k = 1$ for all k, then the same is true for the average:

$$\mathbf{1}^T z_j = \frac{1}{|G_j|} \sum_{k \in G_j} \mathbf{1}^T x_k = \frac{|G_j|}{|G_j|} = 1.$$

(c) The ith entry of group represenative z_j is the fraction of the vectors in group j that have ith entry one. If it is equal to one, all vectors in the group have ith entry one. If it is close to one, most vectors in the group have ith entry one. If it zero, no vectors in the group have ith entry one.

5. *Principal components analysis.* Suppose we are given a set of points $\{x_1, \ldots, x_N\}$ and that this training data has been preprocessed to have zero mean and unit variance in each coordinate. For a given unit vector u, let $f_u(x)$ be the projection of x onto the direction u, *i.e.*,

$$f_u(x) = \operatorname*{argmin}_{v \in V} \|x - v\|_2,$$

where $V = \{tu \mid t \in \mathbf{R}\}$.

Show that the solution z to the problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^N \|x_i - f_u(x_i)\|_2 \\ \text{subject to} & \|u\|_2 = 1 \end{array}$$

is the first principal component of the training data. In other words, the unit vector that minimizes the mean squared error between projected points and original points corresponds to the first principal component.

Note. This problem shows the following result for the case $k = 1$: The subspace spanned by the first k principal components is the k-dimensional subspace that minimizes the sum of squares distance between the original points and their projections onto the subspace.

Solution. We can show that $f_u(x) = u^T x u$; then it is easy to see that

$$\operatorname*{argmin}_{u:\|u\|_2^2=1} \left(\sum_{i=1}^N \|x_i - u^T x_i u\|_2^2 \right) = \operatorname*{argmax}_{u:\|u\|_2^2=1} u^T \left(\sum_{i=1}^N x_i x_i^T \right) u,$$

and the righthand side is the optimization problem defining the first principal component.