MLPR Tutorial Sheet 1

You should attempt the tutorial questions *before* your tutorial.

We strongly recommend you discuss these questions — and the course in general — with your peers. You could work at ML-Base, or arrange to meet up with people from your tutorial group (e.g., through Learn) or other people you have met in lectures.

If you can't do a part, skip it at first and move on! After attempting what you can, try to meet up with friend from the class and pool your understanding. Tutorials run more smoothly if you have agreed with someone what you want to talk about.

Your tutorial session usually won't discuss every part. The point of tutorials isn't to give you the answers: detailed answers will be made available after the week's tutorials, and can be discussed further on Hypothesis. The tutorial sessions are mainly useful for giving you practice at explaining your thinking, and discussing particular points that the group might find helpful.

---

1. **Linear Regression and linear transformations:**

   Alice fits a function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ to a training set of N datapoints $\{\mathbf{x}^{(n)}, y^{(n)}\}$ by least squares. The inputs $\mathbf{x}$ are D-dimensional column vectors. You can assume a unique setting of the weights $\mathbf{w}$ minimizes the square error on the training set.

   Bob has heard that by transforming the inputs $\mathbf{x}$ with a vector-valued function $\boldsymbol{\phi}$, he can fit an alternative function, $g(\mathbf{x}) = \mathbf{v}^\top \boldsymbol{\phi}(\mathbf{x})$, with the same fitting code. He decides to use a linear transformation $\boldsymbol{\phi}(\mathbf{x}) = A\mathbf{x}$, where A is an invertible matrix.

   a) Show that Bob's procedure will fit the same function as Alice's original procedure.

      NB You don't have to do any extensive mathematical manipulation. You also don't need a mathematical expression for the least squares weights. Hint: reason about the sets of functions that Alice and Bob are choosing their functions from.

   b) Could Bob's procedure be better than Alice's if the matrix A is not invertible?

      [If you need a hint, it may help to remind yourself of the discussion involving invertible matrices in the pre-test answers.]

   c) Alice becomes worried about overfitting, adds a regularizer $\lambda \mathbf{w}^\top \mathbf{w}$ to the least-squares error function, and refits the model. Assuming A is invertible, can Bob choose a regularizer so that he will still always obtain the same function as Alice?

   d) **Bonus part:** *Only do this part this week if you have time. Otherwise review it later.* Suppose we wish to find the vector $\mathbf{v}$ that minimizes the function

      $$(\mathbf{y} - \Phi\mathbf{v})^\top (\mathbf{y} - \Phi\mathbf{v}) + \mathbf{v}^\top M \mathbf{v}.$$

      i) Show that $\mathbf{v}^\top M \mathbf{v} = \mathbf{v}^\top (\frac{1}{2}M + \frac{1}{2}M^\top)\mathbf{v}$, and hence that we can assume without loss of generality that M is symmetric.

      ii) Why would we usually choose M to be *positive semi-definite* in a regularizer, meaning that $\mathbf{a}^\top M \mathbf{a} \geq 0$ for all vectors $\mathbf{a}$?

      iii) Assume we can find a factorization $M = AA^\top$. Can we minimize the function above using a standard routine that can minimize $(\mathbf{z} - X\mathbf{w})^\top (\mathbf{z} - X\mathbf{w})$ with respect to $\mathbf{w}$?

2. **Logistic Sigmoids:**

   i) Sketch — with pen and paper — a contour plot of the sigmoidal function

   $$\phi(\mathbf{x}) = \sigma(\mathbf{v}^\top \mathbf{x} + b),$$

   for $\mathbf{v} = [1\ 2]^\top$ and $b = 5$, where $\sigma(a) = 1/(1+\exp(-a))$.

   Indicate the precise location of the $\phi = 0.5$ contour on your sketch, and give at least a rough indication of some other contours. Also mark the vector $\mathbf{v}$ on your diagram, and indicate how its direction is related to the contours.

   Hints: What happens to ϕ as $\mathbf{x}$ moves orthogonal (perpendicular) to $\mathbf{v}$? What happens to ϕ as $\mathbf{x}$ moves parallel to $\mathbf{v}$? To draw the $\phi = 0.5$ contour, it may help to identify special places where it is easy to show that $\phi = 0.5$.

   ii) If $\mathbf{x}$ and $\mathbf{v}$ were three-dimensional, what would the contours of ϕ look like, and how would they relate to $\mathbf{v}$? (A sketch is not expected.)

3. **Radial Basis Functions (RBFs):**

   In this question we form a linear regression model for one-dimensional inputs: $f(x) = \mathbf{w}^\top \boldsymbol{\phi}(x; h)$, where $\boldsymbol{\phi}(x; h)$ evaluates the input at 101 basis functions. The basis functions

   $$\phi_k(x) = e^{-(x-c_k)^2/h^2}$$

   share a common user-specified bandwidth h, while the positions of the centers are set to make the basis functions overlap: $c_k = (k - 51)h/\sqrt{2}$, with $k = 1 \ldots 101$. The free parameters of the model are the bandwidth h and weights $\mathbf{w}$.

   The model is used to fit a dataset with $N = 70$ observations each with inputs $x \in [-1, +1]$. Assume each of the observations has outputs $y \in [-1, +1]$ also. The model is fitted for any particular h by transforming the inputs using that bandwidth into a feature matrix Φ, then minimizing the regularized least squares cost:

   $$C = (\mathbf{y} - \Phi\mathbf{w})^\top (\mathbf{y} - \Phi\mathbf{w}) + 0.1\mathbf{w}^\top \mathbf{w}.$$

   a) Explain why many of the weights will be close to zero when $h = 0.2$, and why even more weights will probably be close to zero when $h = 1.0$.

   b) It is suggested that we could choose h by fitting $\mathbf{w}$ for each h in a grid of values, and pick the h which led to a fit with the smallest cost C. Explain whether this suggestion is a good idea, or whether you would modify the procedure.

   c) Another data set with inputs $x \in [-1, +1]$ arrives, but now you notice that all of the observed outputs are larger, $y \in [1000, 1010]$. What problem would we encounter if we performed linear regression as above to this data? How could this problem be fixed?