6.867 Machine Learning

Problem set 1

Due Friday, September 19, in recitation

Please address all questions and comments about this problem set to `6.867-staff@ai.mit.edu`. You do not need to use MATLAB for this problem set though you can certainly do so. We will provide helpful hints along the way but if you are not familiar with MATLAB and wish to use MATLAB in this problem set, please consult

`http://www.ai.mit.edu/courses/6.867/matlab.html`

and the links therein.

Part 1: Least-Squares Regression

Reference: Lectures 2 and 3, chapters 5-5.3

The goal of this section is to solidify basic concepts in least squares regression. Suppose we have some simple dataset, $\{(x_i, y_i), i = 1, ..., n\}$, where x_i and y_i are real numbers. Our model of how y is related to x is given by

$$\begin{aligned} y &= f(x; w) + e \qquad (1) \\ f(x; w) &= w'\phi(x) \qquad (2) \end{aligned}$$

where $\phi : \mathcal{R} \rightarrow \mathcal{R}^d$ is a specified function (see below) which maps x to a d-dimensional "feature" vector, $\phi(x) = (\phi_1(x), \ldots, \phi_d(x))'$; w is a d-dimensional parameter vector $w = (w_1, \ldots, w_d)'$; e is the prediction error, which we do not model explicitly. We will use w' to denote the transpose of any vector w, as is done in MATLAB. Note that our formulation above does not explicitly include the offset parameter, or w_0, as was done in the lectures. We can incorporate the offset by defining $\phi_1(x) = 1$.

In the following, we wish to determine the least squares optimal parameters or $\hat{w}$. In other words, we minimize the following squared prediction error:

$$J(w) = \frac{1}{n} \sum_i (y_i - f(x_i; w))^2 \qquad (3)$$

By a similar argument as given in Lecture 2, the solution to this problem is

$$\hat{w} = (X'X)^{-1} X'\mathbf{y} \qquad (4)$$

where $X = (\phi(x_1), \ldots, \phi(x_n))'$ is a $n \times d$ matrix whose first row is $\phi_1(x_1), \ldots, \phi_d(x_1)$ and the last row is given by $\phi_1(x_n), \ldots, \phi_d(x_n)$; The output vector $\mathbf{y}$ is defined as $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$.

Note: We assume that the matrix $(X'X)$ is invertible so that the problem is well-posed, i.e. there exists a unique minimizer. This is true when the feature vectors $\phi(x_1), \ldots, \phi(x_n)$ associated with the training examples span the $d-$dimensional feature space. When the feature vectors are long and the number of training points n is small, this is not at all necessarily the case. For example, it cannot be the case whenever $d > n$.

Now, for this estimate $\hat{w}$ the resulting prediction errors $\hat{e}_i = y_i - f(x_i; \hat{w})$ should be "uncorrelated" with the features:

$$\frac{1}{n} \sum_i \hat{e}_i \phi_k(x_i) \quad = \quad 0, \quad k = 1, \ldots, d \tag{5}$$

These conditions are obtained by taking the derivative of $J(w)$ with respect to each w_i, $i = 1, \ldots, d$, and setting them to zero. Note that the prediction error need not be zero mean unless one of the features is a constant, i.e., if say $\phi_1(x) = 1$ for all x, so that

$$\frac{1}{n} \sum_i \hat{e}_i \phi_1(x_i) = \frac{1}{n} \sum_i \hat{e}_i = 0 \tag{6}$$

The error is guaranteed to be "uncorrelated" with only features actually included in the prediction.

You may wonder why we are talking about "correlation" in the first place. Let's explore this a bit further, here and in the problems that follow. For joint samples (u_i, v_i), $i = 1, \ldots, n$, the sample covariance is defined as

$$\tilde{\Sigma}_{u,v} = \frac{1}{n} \sum_i (u_i - \bar{u})(v_i - \bar{v}) \tag{7}$$

where $\bar{u} = (1/n) \sum_i u_i$ is the sample mean of u and similarly for v. The samples are *uncorrelated* if the sample covariance is exactly zero, $\tilde{\Sigma}_{u,v} = 0$. Covariance measures how well one variable (or one set of samples) is linearly predictable from the other.

Problem 1

1. (5pts) Assuming that the first component of the feature vector is a constant, i.e., $\phi_1(x) = 1$, show that the joint "samples" $(\hat{e}_i, \phi_k(x_i))$, $i = 1, \ldots, n$, are indeed uncorrelated for all $k = 1, \ldots, d$ according to our definition above.

2. (5pts) Show that all linear functions of the basis functions, i.e., functions of the form $w'\phi(x)$ for some $w \in \mathcal{R}^d$, are also uncorrelated with the prediction errors $\hat{e}_i$ associated with the least squares optimal parameters $\hat{w}$. In other words, show that $(\hat{e}_i, w'\phi(x_i))$, $i = 1, \ldots, n$, are uncorrelated for any w.

3. (5pts) Yet another way of understanding this result is that if we try to fit a linear function (using the same basis functions) to the prediction errors, we can only get zero. Let $\hat{w}$ and $\hat{e}_i$, $i = 1, \ldots, n$, be defined as above. If we now use $\tilde{y}_i = \hat{e}_i$ as the new target outputs and repeat the parameter estimation step using these new outputs and the same set of basis functions, show that the resulting new least squares parameters are indeed identically zero.

4. (5pts) Suppose we change our feature representation of examples by rescaling the basis functions, i.e., use $\tilde{\phi}(x) = (a_1\phi_1(x), \ldots, a_d\phi_d(x))'$ as the feature vector, where a_i, $i = 1, \ldots, d$ are any non-zero real numbers. Show that the unscaled solution, the function $\hat{w}'\phi(x)$, is still optimal in the sense that $\hat{w}'\phi(x) = \hat{\tilde{w}}'\tilde{\phi}(x)$, where $\hat{\tilde{w}}$ are the least squares optimal parameters for the scaled feature vectors. (*Hint.* use correlation).

5. (Optional) Let's go through a small numerical example to get started with MATLAB. We will use the following data (expressed in MATLAB notation):

   ```
   x = [-2 -1 0 1 2]'; y = [-2 -1 -1 0 1]';
   ```

   (both are column vectors). Let $\phi(x) = (1, x, x^2)'$. To find the least squares parameters, say `wh` in MATLAB, we can construct the `X` matrix simply as

   ```
   X = [ones(size(x)),x,x.^2];
   ```

   where the dot refers to an elementwise operation. Matrix inverse in MATLAB is `inv(A)` for any invertible `A`.

   Find the least squares optimal parameters `wh` in this case. Plot the sample points and the resulting function corresponding to the parameters.

   Verify that the prediction error is indeed uncorrelated with the basis functions.

   Repeat the procedure for $\phi(x) = sin(\pi x)$ (only one basis function). (note: π in MATLAB is simply a constant `pi`). Does the result look reasonable? What should the answer be?

Problem 2

The predictions we make in the regression formulation need not be one dimensional. We can just as easily make predictions that are vector valued. Consider a simple example where the input x takes only binary values $x \in \{0, 1\}$ and $\mathbf{y}$ is a two-dimensional measurement $\mathbf{y} \in \mathcal{R}^2$. Here, the model is

$$\mathbf{y} = \mathbf{f}(x; W) + \mathbf{e} \tag{8}$$

$$\mathbf{f}(x; W) = W'\phi(x) \tag{9}$$

where the feature vector is defined by $\phi(0) = (1,0)'$ and $\phi(1) = (0,1)'$; W is a two-by-two matrix of model parameters; and both the prediction $\mathbf{f}(x;W)$ and the prediction errors $\mathbf{e}_i = \mathbf{y}_i - \mathbf{f}(x_i;W)$ are two-dimensional vectors. We now wish to determine $\hat{W}$ that minimizes the squared error

$$J(W) = \frac{1}{n}\sum_i ||\mathbf{e}_i||^2 = \frac{1}{n}\sum_i \mathbf{e}_i'\mathbf{e}_i \tag{10}$$

1. (10pts) Show that the least-squares estimate of W is

   $$\hat{W} = (X'X)^{-1}X'Y \tag{11}$$

   where $X = (\phi(x_1)\ldots\phi(x_n))'$ and $Y = (\mathbf{y}_1\ldots\mathbf{y}_n)'$. *Hint.* Show that the objective decomposes so that each column of W may be obtained independently; you are essentially solving two separate 1-dimensional regression problems.

   Next, consider the data set:

x	y
0	$(-1,-1)'$
0	$(-1,-2)'$
0	$(-2,-1)'$
1	$(1,1)'$
1	$(1,2)'$
1	$(2,1)'$

2. (5pts) Compute $\hat{W}$. Plot the data points $\mathbf{y}_i$ and the columns of $\hat{W}' = (\hat{\mathbf{y}}_0\ \hat{\mathbf{y}}_1)$ (note the transpose).

3. (5pts) Verify that $\sum_i \hat{\mathbf{e}}_i\phi(x_i)' = 0$ (a 2x2 matrix in this case). What is the interpretation of the columns of $\hat{W}'$?

Part 2: Probabilistic Modeling and Likelihood

Reference: Lecture 3, chapter 4 (up to eq 4.20)

First a bit of background. Suppose we have a probability distribution or density $p(x;\theta)$, where x may be discrete or continuous depending on the problem we are interested in. θ specifies the parameters of this distribution such as the mean and the variance of a one dimensional Gaussian. Different settings of the parameters imply different distributions over x. The available data, when interpreted as samples $x_1,\ldots,x_n$ from one such distribution, should favor one setting of the parameters over another. We need a formal criterion for gauging how well any potential distribution $p(\cdot|\theta)$ "explains" or "fits" the data. Since

$p(x|\theta)$ is the probability of reproducing any observation x, it seems natural to try to maximize this probability. This gives rise to the *Maximum Likelihood* estimation criterion for the parameters θ:

$$\hat{\theta}_{ML} = \operatorname*{argmax}_{\theta} L(x_1, \dots, x_n; \theta) = \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} p(x_i|\theta) \tag{12}$$

where we have assumed that each data point x_i is drawn independently from the same distribution so that the likelihood of the data is $L(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta)$. Likelihood is viewed primarily as a function of the parameters, a function that depends on the data.

The above expression can be quite complicated (depending on the family of distributions we are considering), and make maximization technically challenging. However, any monotonically increasing function of the likelihood will have the same maxima. One such function is *log-likelihood* $\log L(x_1, \dots, x_n; \theta)$; taking the log turns the product into a sum, making derivatives significantly simpler. We will maximize the log-likelihood instead of likelihood.

Problem 4

Let $x \in \{0, 1\}$ denote the result of a coin flip ($x = 0$ for ”tails”, and $x = 1$ for ”heads”). The coin is potentially biased so that ”heads” occurs with probability θ_1. Suppose also that someone else observes the coin flip and reports to you ”heads” or ”tails” (denote this report by y). But this person is unreliable and only reports the result correctly with probability θ_2 (the correctness of the report is independent of the coin toss).

1. (5pts) Write down the joint probability distribution $P(x, y|\theta)$ for all x, y (a 2x2 matrix) as a function of the parameters $\theta = (\theta_1, \theta_2)$.

   Suppose we have access to the following (joint) observations of x and y:

x	y
1	1
1	0
0	0
1	0
1	1
0	0
0	1

2. (10pts) What are the maximum-likelihood (ML) values of θ_1 and θ_2? Provide the details of your derivation as well as the answer. *Hint.* You can first confirm that $P(x, y|\theta) = P(y|x, \theta_2)P(x|\theta_1)$, where the key observation is that the parameters can be separated into the different components. After all the distribution of the coin toss, governed by $P(x|\theta_1)$, is independent of the accuracy of the report, contained in $P(y|x, \theta_2)$. This separation helps you to isolate the estimation of each parameter in the log-likelihood criterion.

3. (10pts) Let $\hat{\theta}_1(x_1,\ldots,x_n)$ be the ML estimator of θ_1 based on the observed data $x_1,\ldots,x_n$, where the data is viewed as independent samples from $P(x|\theta_1)$ for some fixed θ_1. We can try to assess how well the estimator recovers the parameters θ_1. One useful measure is the *bias* of the estimator. This is defined as the expectation $\mathbf{E}\left[\hat{\theta}_1(X_1,\ldots,X_n) - \theta_1\right]$, taken with respect to the true distribution of $X_1,\ldots,X_n$ or $\prod_i P(X_i|\theta_1)$. The bias measures whether the estimator systematically deviates from the true parameters θ_1 that were used to generate the data. An estimator is called *unbiased* if its bias is zero. Show that the ML estimator $\hat{\theta}_1$ is indeed unbiased in this sense.

4. (10pts) We have thus far used only two parameters θ_1 and θ_2 to specify the joint distribution over (x,y). This was possible because of the assumption that the accuracy of the report (whether $y = x$) is independent of the coin toss (what x is). It takes three parameters to specify an unconstrained joint distribution over (x,y). While there are four possible configurations of the variables, there are only three parameters that can be set independently (the fourth one is determined due to normalization, $\sum_{x,y} P(x,y) = 1$). We can parameterize the joint distribution symmetrically in terms of four numbers $P(x,y) = \theta_{x,y}$, that sum to one $\sum_{x,y}\theta_{x,y} = 1$. When we estimate the maximum likelihood joint distribution, we estimate the ML setting of the parameters $\hat{\theta}_{x,y}$. What is the maximum likelihood estimate of $P(x,y)$ in this case? Which model has the higher log-likelihood?

5. (Optional) Show that the ML parameters $\hat{\theta}_{x,y}$ are unbiased estimates of $\theta_{x,y}$.

6. (Optional) Suppose we are not sure which model is correct. Can you extend the leave-one-out cross-validation procedure described in the linear regression context to our setting here? Which model would the resulting cross-validation criterion choose in this case?

Problem 5

Consider a bivariate Gaussian distribution $\mathbf{x} = (x_1, x_2)' \sim N(\mu, \Sigma)$ with probability density

$$p(\mathbf{x}; \mu, \Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\} \tag{13}$$

where $\mu = E\{\mathbf{x}\}$ is the two-dimensional mean vector and $\Sigma = E\{(\mathbf{x}-\mu)(\mathbf{x}-\mu)'\}$ is the two-by-two covariance matrix ($|\Sigma|$ is the determinant of the covariance matrix). The Gaussian is fully specified by the parameters (μ, Σ).

1. (10pts) Given a collection of independent samples $\mathbf{x}_i$, $i = 1,\ldots,n$, we wish to estimate the model parameters (μ, Σ). The maximum-likelihood estimates are chosen so as to

maximize the log-likelihood

$$\begin{aligned} J(\mu, \Sigma) &= \log p(\mathbf{x}_1, \ldots, \mathbf{x}_n; \mu, \Sigma) & (14) \\ &= \sum_i \log p(\mathbf{x}_i; \mu, \Sigma) & (15) \end{aligned}$$

Show that the ML estimates based on data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are given by the sample mean and sample covariance:

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_i \mathbf{x}_i & (16) \\ \hat{\Sigma} &= \frac{1}{n} \sum_i (\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})' & (17) \end{aligned}$$

Hints. Start with the mean estimate. Express the Gaussian distribution in terms of the inverse covariance matrix $A = \Sigma^{-1}$ and use the following matrix derivatives

$$\frac{d}{dA} (\mathbf{x} - \mu)' A (\mathbf{x} - \mu) = (\mathbf{x} - \mu)(\mathbf{x} - \mu)' \qquad \frac{d}{dA} \log |A| = A^{-1} \qquad (18)$$

2. (10pts) The bi-variate Gaussian distribution allows the two variables to be dependent on each other (the values of the variables co-vary). This dependence is fully described by the covariance matrix. In light of problem 1 we suspect that this dependence is captured by linearly predicting one from the other. Suppose we have access to x_1 part of the samples from a Gaussian model (μ, Σ) and wish to use them to estimate x_2. Derive the least squares optimal estimate $\hat{x}_2(x_1)$ that minimizes the expected squared error $E\{(x_2 - \hat{x}_2(x_1))^2\}$ (the exectation is over samples $(x_1, x_2)' \sim N(\mu, \Sigma)$). *Hint.* Use the fact that the best estimate is of the form $\hat{x}_2(x_1) = E\{x_2 | x_1\}$, as discussed in the lecture.