Machine Learning Worksheet 2

Linear Regression

1 Parameter Estimation

Consider n samples $x_1, \ldots, x_n$ drawn independently and identically (i.i.d.) from a given distribution $P(X|\theta)$. This distribution is usually parametrized (e.g. one parameter representing its mean, one its variance, etc.); these parameters are denoted by θ. One wants to find accurate estimates for these parameters using the n samples only. *Maximum Likelihood Estimation* (MLE) finds estimates for the various parameters at hand by maximizing the likelihood $P(x_1, x_2, \ldots, x_n|\theta) = \Pi_{i=1}^{n} P(x_i|\theta)$. (i.e. the probability of observing the n samples at hand). Note that usually one considers the *log likelihood*, $\log P(x_1, \ldots, x_n|\theta))$.

1.1 Coins

Let X be a Bernoulli random variable. The Bernoulli distribution is only parametrized by one parameter, $\theta = P(X = 1)$.

Problem 1. For n i.i.d. observations of X determine the MLE for θ. You might want to use $P(X = x|\theta) = \theta^x(1-\theta)^{1-x}$.

Now we look at slightly more complex distribution, the binomial distribution.

Problem 2. $\star$ Consider a binomial random variable X, with prior distribution for μ given by the beta distribution, and suppose we have observed m occurences of $X = 1$ and l occurences of $X = 0$. Show that the posterior *mean* value of μ lies between the prior mean of μ and the maximum likelihood estimate for μ. To do this, show that the posterior mean can be written as λ times the prior mean plus $(1-\lambda)$ times the maximum likelihood estimate, with $0 \leq \lambda \leq 1$. This illustrates the concept of the posterior mean being a compromise between the prior distribution and the maximum likelihood solution.

Note: The binomial distribution is defined as follows:

$$p(x = m|N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

1.2 Poisson distribution

Let X be Poisson distributed.

Problem 3. Again, for n i.i.d. samples from X, determine the maximum likelihood estimate for λ. Show that this estimate is unbiased!

2 Weighted Linear Regression

Consider a linear regression problem in which we want to "weight" different training examples differently. Specifically, suppose we want to minimize

$$E(\boldsymbol{w}) = \frac{1}{2}\sum^{N} \theta_n \left(t_n - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n)\right)^2$$

Problem 4. We already worked out what happens for the case where all the weights θ_n are the same. In this problem, we will generalize some of those ideas to the weighted setting, and also implement the locally weighted linear regression algorithm.

1. Show that $E(\boldsymbol{w})$ can also be written

$$E(\boldsymbol{w}) = (\boldsymbol{T} - \boldsymbol{\Phi w})^T \boldsymbol{\Theta} (\boldsymbol{T} - \boldsymbol{\Phi w}) \tag{1}$$

   for an appropriate diagonal matrix $\boldsymbol{\Theta}$, and where $\boldsymbol{\Phi}$ and $\boldsymbol{T}$ are as defined in class. State clearly what $\boldsymbol{\Theta}$ is.

2. Now let all the θ_n equal 1. By differentiating Eq. 1 with respect to $\mathbf{w}$, derive the normal equations for the least squares problem, as given in class.

3. Generalize the normal equations to the case of arbitrary θ_ns.

4. Suppose we have a training set $(\boldsymbol{x}_n, t_n); n = 1, \ldots, N$ of N independent examples, but in which the t_n were observed with differing variances. Specifically, suppose that

$$p(t_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(t_n|\boldsymbol{w}^T \Phi(\boldsymbol{x}_n), \sigma_n^2)$$

   where the σ_n are fixed, known, constants. Show that finding the maximum likelihood estimate of $\boldsymbol{w}$ reduces to solving a weighted linear regression problem. State clearly what the θ_n are in terms of the σ_n.

3 Basisfunctions

Problem 5. Show that the tanh function and the logistic sigmoid function are related by

$$\tanh(x) = 2\sigma(2x) - 1$$

Thus, show that a general linear combination of logistic sigmoid functions of the form

$$y(x, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M} w_j \sigma\left(\frac{x - \mu_j}{s}\right)$$

is equivalent to a linear combination of tanh functions of the form

$$y(x, \boldsymbol{u}) = u_0 + \sum_{j=1}^{M} u_j \tanh\left(\frac{x - \mu_j}{2s}\right)$$

and find expressions to relate the new parameters $\{u_0, \ldots, u_M\}$ to the original parameters $\{w_0, \ldots, w_M\}$.

Problem 6. Show that that the least square solution for linear regression corresponds to an orthogonal projection of the vector $\boldsymbol{T}$ onto the manifold S as shown in Figure 1. There, the subspace S is spanned by the basis functions $\phi_j(\boldsymbol{x})$ in which each basis function is viewed as a vector $\boldsymbol{\varphi}_j$ of length N with elements $\phi_j(\boldsymbol{x}_n)$. (Hint: You might want consider what $\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T$ resembles, e.g. how does it relate to the maximum likelihood solution for linear regression.)

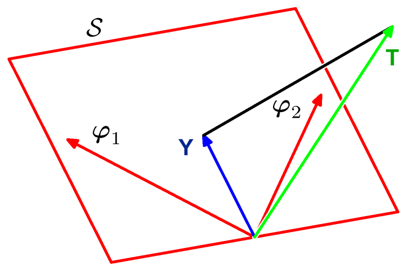

Figure 1: The projection property of $\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T$.

4 Bayesian Linear Regression

Problem 7. $\star$ We have seen that, as the size of a data set increases, the uncertainty associated with the posterior distribution over model parameters decreases (see worksheet 1). Prove the following matrix identity

$$(\boldsymbol{M} + \boldsymbol{v}\boldsymbol{v}^T)^{-1} = \boldsymbol{M}^{-1} - \frac{(\boldsymbol{M}^{-1}\boldsymbol{v})(\boldsymbol{v}^T\boldsymbol{M}^{-1})}{1 + \boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}}$$

and, using it, show that the uncertainty $\sigma_N^2(\boldsymbol{x})$ associated with the bayesian linear regression function given by eq. (33) in the slides satisfies

$$\sigma_{N+1}^2(\boldsymbol{x}) \leq \sigma_N^2(\boldsymbol{x})$$