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 $\theta$. 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 $P(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} P(x_i|\theta)$. (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 $\theta$. 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.** Consider a binomial random variable $X$, with prior distribution for $\mu$ 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 $\mu$ lies between the prior mean of $\mu$ and the maximum likelihood estimate for $\mu$. To do this, show that the posterior mean can be written as $\lambda$ 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 $\lambda$. 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

\[ E(w) = \frac{1}{2} \sum_{n=1}^{N} \theta_n (t_n - w^T \phi(x_n))^2 \]

Problem 4. We already worked out what happens for the case where all the weights \( \theta_n \) 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 \( E(w) \) can also be written

\[ E(w) = (T - \Phi w)^T \Theta (T - \Phi w) \] (1)

for an appropriate diagonal matrix \( \Theta \), and where \( \Phi \) and \( T \) are as defined in class. State clearly what \( \Theta \) is.

2. Now let all the \( \theta_n \) equal 1. By differentiating Eq. 1 with respect to \( w \), derive the normal equations for the least squares problem, as given in class.

3. Generalize the normal equations to the case of arbitrary \( \theta_n \)s.

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

\[ p(t_n|x_n, w) = \mathcal{N}(t_n|w^T \Phi(x_n), \sigma^2_n) \]

where the \( \sigma_n \) are fixed, known, constants. Show that finding the maximum likelihood estimate of \( w \) reduces to solving a weighted linear regression problem. State clearly what the \( \theta_n \) are in terms of the \( \sigma_n \).

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

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

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

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

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

![Figure 1: The projection property of $\Phi(\Phi^T\Phi)^{-1}\Phi^T$.](image)

4 Bayesian Linear Regression

**Problem 7.** 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

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

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

$$ \sigma^2_{N+1}(\mathbf{x}) \leq \sigma^2_N(\mathbf{x}) $$