

ICS273A: HW2

Due: Feb 4

Problem I: Weighted linear regression. Let the cost function for weighted linear regression be

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m w^{(i)} (y^{(i)} - \theta^T x^{(i)})^2$$

- A Write the cost function using matrix notation, where θ is the linear model, X is the design matrix, \bar{y} is a m -length vector of $y^{(i)}$ values, and W is a diagonal matrix. State what the diagonal entries of W are.
- B Derive a closed form solution to the above cost function using a weighted form of the normal equations (recall the original equation was $X^T(\bar{y} - X\theta) = 0$, with a solution of $\theta = (X^T X)^{-1} X^T \bar{y}$).
- C Suppose we have a training set of m training pairs $\{(x^{(i)}, y^{(i)}) : i = 1 \dots m\}$, but each label $y^{(i)}$ was observed with a fixed, known variance $\sigma^{(i)}$. We can write

$$p(y^{(i)} | x^{(i)}, \theta) = \frac{1}{\sqrt{2\pi}\sigma^{(i)}} \exp \frac{-(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}$$

Show that the above cost function finds the θ that maximizes the log likelihood of this probabilistic model. Describe how $\sigma^{(i)}$ relates to $w^{(i)}$.

Problem II: Exponential family & generalized linear models. A probability distribution in the exponential family takes on the following form:

$$\begin{aligned} p(x|\eta) &= h(x) \exp\{\eta^T T(x) - A(\eta)\} \\ &= h(x) \exp\left\{\sum_i \eta_i T_i(x) - A(\eta)\right\} \end{aligned}$$

where h is called the reference function, T is the sufficient statistic, η is the natural parameter, and $A(\eta)$ is a normalization constant.

A Show that the following distributions are in the exponential family, defining the T, A, and h functions in each case. Make sure to write $A(\eta)$ as a function of the natural parameters η .

1. A unit variance gaussian random parameterized by μ . Note: you can do this using a one dimensional parameter vector η .

$$p(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2}\right\}$$

2. A poisson random variable parameterized by λ

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

3. A multinomial random variable, which can be thought of as the outcome of a biased k -sided die, where θ_i is the probability of rolling an i . Write the roll x as k -length vector of all zeros with a single one for the rolled index i .

$$p(x|\theta) = \theta_1^{x_1} \theta_2^{x_2} \dots \theta_k^{x_k}$$

Use the fact that $\theta_k = 1 - \sum_{i=1}^{k-1} \theta_i$, and similarly $x_k = 1 - \sum_{i=1}^{k-1} x_i$. Express the distribution as a $(k-1)$ dimensional parameter vector η .

B The function $A(\eta)$ can be used to calculate *moments* of the random variable x . Show the following equations hold

$$A(\eta) = \log \int h(x) \exp\{\eta^T T(x)\} dx \quad (1)$$

$$\frac{\partial A}{\partial \eta_i} = E[T_i(x)] \quad (2)$$

$$\frac{\partial^2 A}{\partial \eta_i \partial \eta_j} = \text{cov}(T_i(x), T_j(x)) \quad (3)$$

Hints: For (1), use the fact that $\int p(x|\eta) dx = 1$. For (2), assume the derivative operator can pass inside the integral from (1) and use the fact that $\exp A(\eta) = \int h(x) \exp\{\eta^T T(x)\} dx$. For (3), recall $\text{cov}(A, B) = E[AB] - E[A]E[B]$.

Sanity Check for Part A: You can apply Eq (2) to the three distributions from part (A) to calculate the means. If you have correctly defined $A(\eta)$, you should recover the original parameters μ , λ , and θ .

C Generalized linear models (GLM). Assume we want to predict y given x using a probability distribution from the exponential family. For simplicity, assume both x and the canonical parameters η are one dimensional, and the sufficient statistics $T(y) = y$. We can write $p(y|\eta) = h(y)\exp(\eta y - A(\eta))$. We can write the log-likelihood for a set of training points as

$$l(\theta) = \sum_i \log p(y^{(i)}|\eta^{(i)})$$

where $\eta = \theta x$ for a GLM. Show that

$$\frac{\partial^2 l}{\partial \theta \partial \theta} \leq 0$$

Hint: Use both results from Part B, and the fact that the variance of a random variable must always be positive. This proves that, in the one dimensional case, GLMs (including linear and logistic regression) are concave (or alternatively, the negative log likelihood is convex). This also holds in the vector-valued case.

Problem III K-way linear discriminant analysis. Assume we are given a training set of examples where $x^{(i)} \in R^n$ and $y^{(i)} \in \{1 \dots k\}$. We will model y as multinomial random variable, and will model x conditioned on y as a multivariate gaussian with a single covariance matrix Σ for all k classes.

$$p(y|\theta) = \theta_1^{y_1} \theta_2^{y_2} \dots \theta_k^{y_k} \quad (4)$$

$$p(x|y_i = 1) = \frac{1}{(2\pi)^{(n/2)}|\Sigma|^{1/2}} \exp -\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i) \quad (5)$$

A Show that the posterior for y can be written as the soft-max function below. You may have to redefine x to include a constant term $x_0 = 1$.

$$p(y = i|x, \theta, \Sigma, \{\mu_j\}) = \frac{\exp(w_i^T x)}{\sum_j \exp(w_j^T x)}$$

B Show that for $k = 2$, the posterior takes the form of a logistic function

$$p(y = 1|x, \theta, \Sigma, \mu_1, \mu_2) = \frac{1}{1 + \exp(-w^T x)}$$

Problem IV [MATLAB] Logistic regression. Download hw2train.dat from the website. Each datapoint consists of a two-dimensional input feature and an associated binary label.

A Plot the data, using 0s and Xs for the two classes. The plots in the following parts should be plotted on top of this. You can use the command `hold on` to do this.

- B Fit a generative model to the data, using Gaussian class-conditional densities with equal covariance matrices. Calculate the posterior probability of class 1, and plot the line where the probability is equal to 0.5
- C Write a program to fit a logistic regression model using the IRLS algorithm (remember to include the intercept term). Recall this method uses the Newton-Raphson method to optimize the log likelihood of the training data. Plot the line where the logistic function is equal to 0.5.
- D Write a program to fit a logistic regression model using stochastic gradient ascent. Plot the line where the logistic function is equal to 0.5.
- E Fit a linear regression to the problem, treating the class labels as real values 0 and 1. Use the normal equations to solve for the linear predictor. Plot the line where the predictor function is equal to 0.5.
- F The data set hw2test.dat is a separate data set generated from the same source. Test your fits from parts (b), (c), (d), and (e) by computing the fraction of miss-classified test points.