CS281A - Problem Set 2

Andrea Bajcsy

September 19, 2016

Problem 2.1.

(a) We can formulate the polynomial regression problem as a form of linear prediction by solving the general linear model equation $X\alpha = y$ where:

$$X = \begin{bmatrix} 1 & t_1 & t_1^2 & ... & t_1^D \\ 1 & t_2 & t_2^2 & ... & t_2^D \\ 1 & t_3 & t_3^2 & ... & t_3^D \\ ... & & & & \\ 1 & t_n & t_n^2 & ... & t_n^D \end{bmatrix} \alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ ... \\ \alpha_D \end{bmatrix} y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ ... \\ y_n \end{bmatrix}$$

(b) Figure 1 shows a plot of the mean-squared error R(D) vs. Degree $D \in 1, 2, ...n-1$ when using the data in y.dat and t.dat. See back for code that performs least-squares fit of a polynomial of degree D.

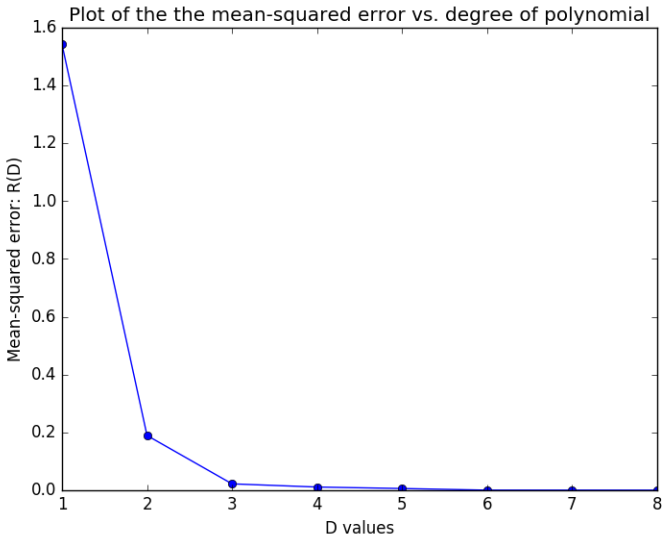

Figure 1: D vs. R(D)

(c) With the degree n-1 fit, we get zero mean-squared error since the function fits exactly to every data point. If we try and fit a polynomial of degree n, we will be solving $X\alpha = y$, where

$X^{n\ x\ n+1}$.

$$\begin{bmatrix} 1 & t_1 & t_1^2 & ... & t_1^n \\ 1 & t_2 & t_2^2 & ... & t_2^n \\ 1 & t_3 & t_3^2 & ... & t_3^n \\ ... & & & & \\ 1 & t_n & t_n^2 & ... & t_n^n \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ ... \\ \alpha_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ ... \\ y_n \end{bmatrix}$$

Using ordinary least-squares, we would solve for $\alpha = (X^T X)^{-1} X^T y$. However, we see there are more columns than rows, so the matrix $X^T X^{-1}$ will not be invertible and we will not be able to obtain a solution.

(d) Figure 2 shows a plot of the degree $D \in 1, 2, ...n-1$ versus the mean-squared error R(D) and $\tilde{R}$ when using the data in y.dat, yfresh.dat, and t.dat. Since the model was trained on y.dat, then it will approximate yfresh.dat with greater error than the data it was trained on since it cannot be a perfect estimator. Note that the error appears to plateu for the same values of D with yfresh.dat or y.dat but at a higher error value when using yfresh.dat. We would want to choose the degree D where the error is at a minimum to ensure that we are not be overfitting the data.

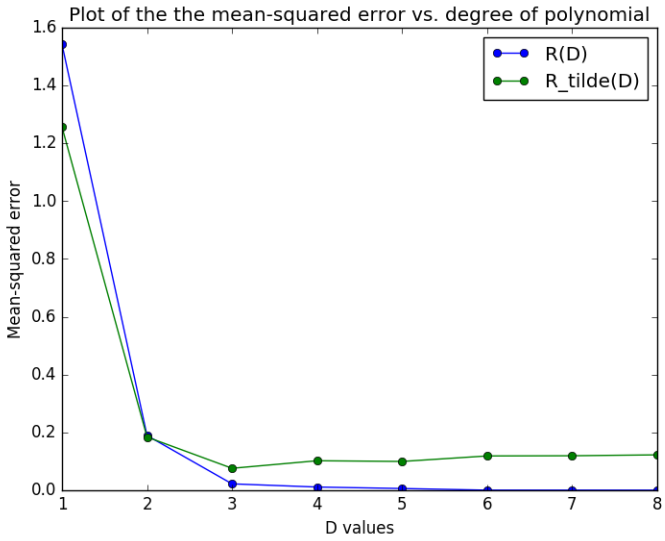

Figure 2: D vs. R(D) and $\tilde{R}$

(e) Figure 3 shows a plot of the degree $D \in 2, ...9$ versus the mean-squared error $\tilde{R}$ and F(D) when using the data in y.dat, yfresh.dat, and t.dat. The minimizing arguments of the two functions are related in that F(D) adds an extra cost based on the number of degrees that we are fitting. The lowest error for R(D) and $\tilde{R}(D)$ occurs when D=3, after which we see a penalization of greater D's in F(D) and higher error rates in $\tilde{R}(D)$.

Problem 2.2.

(a) Prove that A is a convex function.

Proof. By definition,

$$A(\eta) = log(\int_\gamma h(y) e^{\eta y}\, dy) \quad , \quad p_\eta(y) = h(y) e^{\eta y - A(\eta)}$$

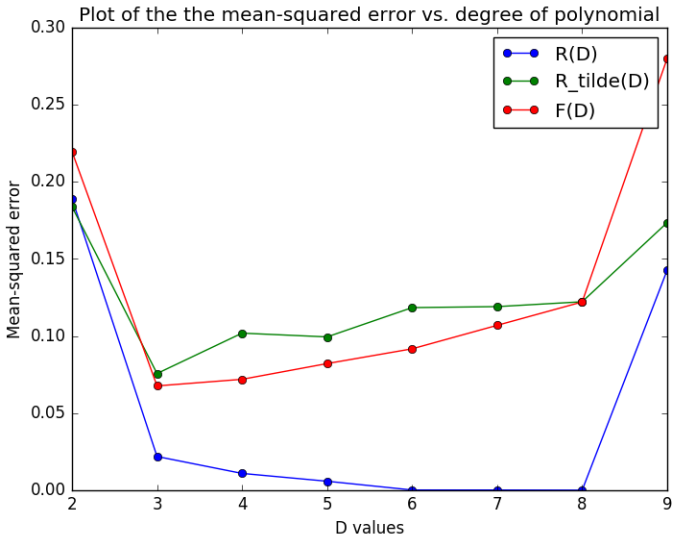

Figure 3: D vs. $\tilde{R}$ and F(D)

To prove convexity, we want to take the second derivative. Let:

$$B(\eta) = \int_{\gamma} h(y)e^{\eta y}\, dy$$

Then the first derivative we get:

$$\frac{\partial A(\eta)}{\partial \eta} = \left(\frac{1}{B(\eta)}\right)\left(\frac{\partial B(\eta)}{\partial \eta}\right) = \frac{\int_{\gamma} h(y)e^{\eta y} y\, dy}{\int_{\gamma} h(y)e^{\eta y}\, dy} = \frac{\int_{\gamma} h(y)e^{\eta y - A(\eta)} y\, dy}{\int_{\gamma} h(y)e^{\eta y - A(\eta)}\, dy} = E_{p_\eta}[y]$$

Taking the second derivative we have:

$$\frac{\partial}{\partial \eta}\frac{B'(\eta)}{B(\eta)} = \frac{\partial}{\partial \eta}\left(B'(\eta)\frac{1}{B(\eta)}\right) = \frac{B''(\eta)}{B(\eta)} - \frac{(B'(\eta))^2}{B(\eta)^2}$$

$$= \frac{\int_{\gamma} h(y)e^{\eta y} y^2\, dy}{\int_{\gamma} h(y)e^{\eta y}\, dy} \; - \; (E_{p_\eta}[y])^2 = \frac{\int_{\gamma} h(y)e^{\eta y - A(\eta)} y\, dy}{\int_{\gamma} h(y)e^{\eta y - A(\eta)}\, dy} \; - \; (E_{p_\eta}[y])^2$$

$$= E_{p_\eta}[y^2] \; - \; (E_{p_\eta}[y])^2 = Var_{p_\eta}[y] \; \succeq \; 0$$

Since Var_{p_η} is positive definite, we have shown that $A(\eta)$ is convex. □

(b) Express KL divergance in terms of $A(\eta)$ and $A'(\eta)$.

$$D(p_\eta || p_\eta) = E_\eta\left(log(\frac{h(y)e^{\eta y - A(\eta)}}{h(y)e^{\tilde{\eta} y - A(\tilde{\eta})}})\right)$$

$$= \int_y log\left(e^{(\eta-\tilde{n})y-(A(\eta)-A(\tilde{\eta}))}p_\eta(y)\right)\, dy$$

$$= \int_y \left((\eta - \tilde{n})y - (A(\eta) - A(\tilde{\eta}))\right) h(y)e^{\eta y - A(\eta)}\, dy$$

$$= (\eta - \tilde{n}) \int_y h(y) e^{\eta y - A(\eta)} y \, dy - (A(\eta) - A(\tilde{\eta})) \int_y h(y) e^{\eta y - A(\eta)} \, dy$$

$$= (\eta - \tilde{n}) A' - A(\eta) + A(\tilde{\eta})$$

Since $A = \int_y h(y) e^{\eta y - A(\eta)} y$ and $\int_y p_\eta(y) dy = 1$ by definition.

(i) Bernoulli random variable:

$$p_\eta(y) = \eta^y (1-\eta)^{1-y}, y \in 0, 1, n \in (0, 1)$$

$$= e^{ylog(\eta) + (1-y)log(1-\eta)} = e^{ylog(\frac{\eta}{1-\eta}) - log(1 + e^{\frac{\eta}{1-\eta}})}$$

Thus, we have $A(\eta) = log(1 + e^\eta)$ and $A^*(t) = sup_{\eta \in R} \ \{\eta t - log(1 + e^\eta\}$. We now take the gradient of A^* with respect to η, set this to 0 in order to solve the optimization problem, and then solve for η in terms of t.

$$\nabla_\eta A^*(t) = t - \frac{e^\eta}{1 + e^\eta}$$

$$0 = t - \frac{e^\eta}{1 + e^\eta} \implies t = \frac{e^\eta}{1 + e^\eta}$$

$$\frac{1}{t} = \frac{1 + e^\eta}{e^\eta} = \frac{1}{e^\eta} + 1 \implies \frac{1}{t} - 1 = \frac{1}{e^\eta}$$

$$e^\eta = \frac{1}{\frac{1}{t} - 1} \implies \eta = log(\frac{1}{\frac{1}{t} - 1})$$

$$\eta = -log(\frac{1}{t} - 1))$$

Substituting this back into our equation, we get:

$$A^*(t) = -tlog(\frac{1}{t} - 1) - log(1 + e^{-log(\frac{1}{t} - 1)}) = -tlog(\frac{1}{t} - 1) + log(1 - t)$$

$$= tlog(t) - tlog(1 - t) + log(1 - t) = tlog(t) + (1 - t)log(1 - t)$$

(ii) Gaussian random variable (based on notes from class):

$$p_\eta(y) = \frac{e^{\frac{-y^2}{2}}}{\sqrt{2\pi}} e^{yn - \frac{\eta^2}{2}}$$

Thus, we have $A(\eta) = \frac{\eta^2}{2}$ and $A^*(t) = sup_{\eta \in R} \ \left\{\eta t - \frac{\eta^2}{2}\right\}$.

$$\nabla_\eta A^*(t) = t - n$$

$$0 = t - n \implies t = n$$

Substituting this back into our equation, we get:

$$A^*(t) = t^2 - \frac{t^2}{2} = \frac{t^2}{2}$$

(iii) Poisson random variable:

$$p_\eta(y) = \frac{1}{y!} e^{y\eta - e^\eta}$$

Thus, we have $A(\eta) = e^\eta$ and $A^*(t) = sup_{\eta \in R}\ \{\eta t - e^\eta\}$.

$$\nabla_\eta A^*(t) = t - e^\eta$$
$$0 = t - e^\eta \implies log(t) = n$$

Substituting this back into our equation, we get:

$$A^*(t) = tlog(t)e^{log(t)} = tlog(t) - t = t(log(t) - 1)$$

(d) Prove that A^* is always a convex function.

Proof. If A^* is always a convex function, then it must satisfy the definition of convexity:

$$A^*(\alpha t_1 + (1-\alpha)t_2) \leq \alpha A^*(t_1) + (1-\alpha)A^*(t_2)$$

We know that A^* is defined by:

$$A^*(\alpha t_1 + (1-\alpha)t_2) = sup_\eta \{\eta(\alpha t_1 + (1-\alpha)t_2) - A(\eta)\}$$
$$= sup_\eta \{\eta(\alpha t_1 + (1-\alpha)t_2) - \alpha A(\eta) - (1-\alpha)A(\eta)\}$$
$$= sup_\eta \{\alpha(\eta t_1 - A(\eta)) + (1-\alpha)(\eta t_2 - A(\eta))\}$$

Let $h(\eta) = \eta t_1 - A(\eta)$ and $k(\eta) = \eta t_2 - A(\eta)$. We know that:

$$sup_\eta(h(\eta) + k(\eta)) \leq sup_\eta(h(\eta)) + sup_\eta(k(\eta))$$

By this property and after resubstituting, we have shown:

$$A^*(\alpha t_1 + (1-\alpha)t_2) \leq \alpha A^*(t_1) + (1-\alpha)A^*(t_2)$$

and that A^* is always a convex function.

□

Problem 2.3.

(a) By definition, we have likelihood of η as:

$$l(\eta; y_1, ...y_n) = log(p(y_1, ..., y_n|\eta)) = log(h(y_1, ...y_n)) + \eta^T \sum_{i=1}^{n} y_i - nA(\eta)$$

We differentiate and solve for $\hat{\eta}$ (assuming the inverse function exists under suitable regularity conditions):

$$\frac{\partial}{\partial \eta} l(\eta; y_1, ...y_n) = \sum_{i=1}^{n} y_i - n\frac{\partial}{\partial \eta} A(\eta)$$
$$\frac{\partial}{\partial \eta} A(\eta) = \frac{\sum_{i=1}^n y_i}{n} \implies \hat{\eta} = (A'^{-1})(\frac{\sum_{i=1}^n y_i}{n})$$

(b) Closed-form estimates for MLE in Poisson, Bernoulli, Gaussian models:

Poisson

$$\frac{\partial}{\partial \eta} A(\eta) = e^\eta = \frac{\sum_{i=1}^n y_i}{n}$$

$$\hat{\eta} = log(\frac{\sum_{i=1}^{n} y_i}{n})$$

Bernoulli (where $\frac{\sum_{i=1}^{n} y_i}{n} \neq 1$)

$$\frac{\partial}{\partial \eta} A(\eta) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\hat{\eta} = log(\frac{\frac{\sum_{i=1}^{n} y_i}{n}}{1 - \frac{\sum_{i=1}^{n} y_i}{n}})$$

Gaussian

$$\frac{\partial}{\partial \eta} A(\eta) = \eta = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\hat{\eta} = \frac{\sum_{i=1}^{n} y_i}{n}$$

(c)

We know that $E(\bar{y}) = A'(\eta^*)$ and by definition of MLE with respect to η, we have:

$$max_{\eta} \left\{ \eta \sum_{i=1}^{n} y_i - nA(\eta) \right\} = max_{\eta} \left\{ \eta \bar{y} - A(\eta) \right\} = min_{\eta} \left\{ -\eta \bar{y} + A(\eta) \right\}$$

Thus, we know that as $n \to \infty$, then $\bar{y} \to A'(\eta^*)$. Additionally, we can add or subtract any terms to this equation that do not depend on η, since they are just constants. We can substitute this into the above equation to get:

$$\hat{\eta} = min_{\eta} \left\{ -\eta A'(\eta^*) + A(\eta) \right\} = min_{\eta} \left\{ -\eta A'(\eta^*) + A(\eta) + \eta^* A'(\eta^*) - A(\eta^*) \right\}$$

After some rearranging of terms, we get the final equation:

$$\hat{\eta} = min_{\eta} \left\{ A(\eta) - A(\eta^*) - A'(\eta^*)(\eta - \eta^*) \right\}$$

(d)

Assume A is strictly convex and that η^* is the true parameter. We can look at the objective function that we are trying to minimize to determine how to bound $\eta \neq \eta^*$ such that:

$$A(\eta) > A(\eta) - A(\eta^*) - A'(\eta^*)(\eta - \eta^*) = 0 = A(\eta^*) - A(\eta^* - A'(\eta^*)(\eta^* - \eta^*)$$

Notice that this implies that the minimizing value of η must be η^* as n approaches infinity and for A strictly convex in order for this to hold. Thus, $\hat{\eta} = \eta^*$.

Problem 2.4.

(a) Based on the definition of MLE in GLM as well as stochastic gradient from the course reader, the update step can be written with μ as the canoniucal response function as:

$$\tilde{\theta}^{t+1} = \hat{\theta}^t + p(y_n - \mu_n^t)x_n$$

(b) Explicit updates for Possion and Logistic cases:

Poisson

$$A(t) = e^t \implies \tilde{\theta}^{t+1} = \hat{\theta}^t + p(y_n - e^{x_n^T \theta^t})x_n$$

Logistic

$$A(t) = log(1 + e^t) \implies \tilde{\theta}^{t+1} = \hat{\theta}^t + p(y_n - \frac{1}{1 + e^{x_I^T \theta^t}})x_n$$

(c) Figure 4 shows a histogram plot of the probabilities $P[y_i|x_i; \hat{\theta}]$ based on fitted vector $\hat{\theta}$ and files Xone.dat and yone.dat.

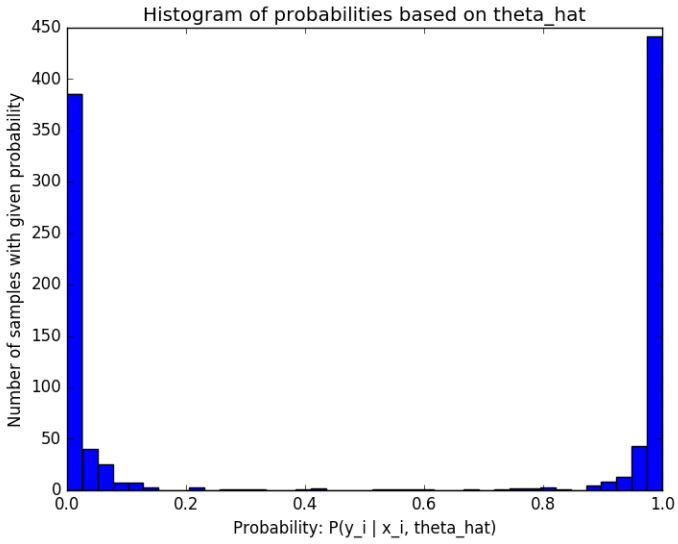

Figure 4: Histogram of probabilities using Xone.dat and yone.dat

(d) Figure 5 shows a histogram plot of the probabilities $P[y_i|x_i; \hat{\theta}]$ based on fitted vector $\hat{\theta}$ and files Xtwo.dat and ytwo.dat.The differences in Figure 4 and 5 suggest that it is easier to separate the data in Figure 4 into two clusters (leading to more accurate fit) than the data in Figure 5 which is all centered around 0.5. It also suggests that the model is not learning anything new with the second dataset becuase most data points are assigned with probability=0.5, which with binary labels is what we would expect at random.

(e) See Figure 6 and 7 for a visualization of a fitted 2-component GMM for Xone.dat and Xtwo.dat using labels in yone.dat and ytwo.dat.

(f) In part (e), the fitted mean vectors and the $\hat{\theta}$:

$$\mu_{Xone1label} = \begin{bmatrix} 3.03708769 \\ -3.03958106 \end{bmatrix} \mu_{Xone0label} = \begin{bmatrix} 2.9674205 \\ 3.02468416 \end{bmatrix}$$

$$\mu_{Xtwo1label} = \begin{bmatrix} 0.03544899 \\ -0.09377857 \end{bmatrix} \mu_{Xtwo0label} = \begin{bmatrix} 0.01322437 \\ -0.00039517 \end{bmatrix}$$

$$\hat{\theta}_{LogXone} = \begin{bmatrix} 0.03971547 \\ -1.69052142 \end{bmatrix} \hat{\theta}_{LogXtwo} = \begin{bmatrix} -0.01176046 \\ -0.02374169 \end{bmatrix}$$

If for each data matrix you take the line perpendicular to the vector formed by the difference of the two means, then the $\hat{\theta}$ vectors will be perpendicular to this line. As seen in Figures 6, 7, this line attempts to partition the data based on the given labels, and the $\hat{\theta}$ is it's normal vector.

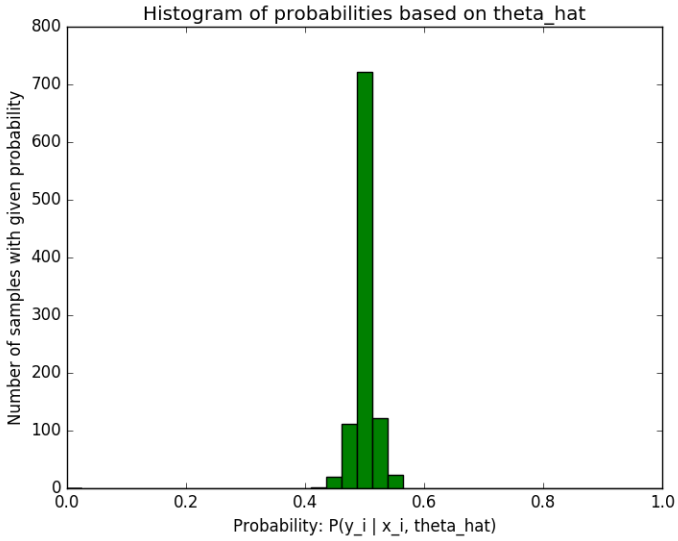

Figure 5: Histogram of probabilities using Xtwo.dat and ytwo.dat

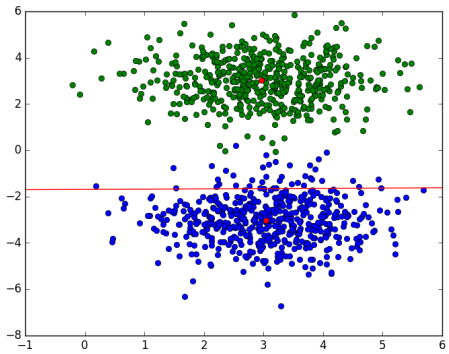

Figure 6: Results of fitting a 2-component GMM to Xone.dat using labels in yone.dat. Red dots indicate the means of the clusters.

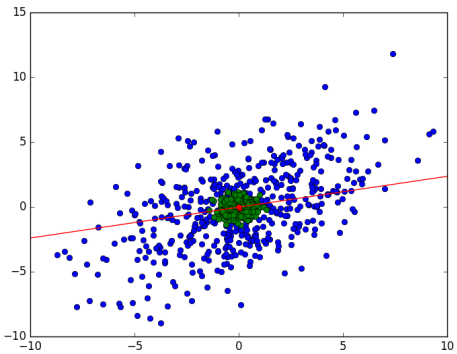

Figure 7: Results of fitting a 2-component GMM to Xtwo.dat using labels in ytwo.dat. Red dots indicate the means of the clusters.