

CS281A - Problem Set 2

Andrea Bajcsy

September 16, 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:

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

(b) Figure 1 shows a plot of the mean-squared error $R(D)$ vs. Degree $D \in 1, 2, \dots, n-1$ 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 .

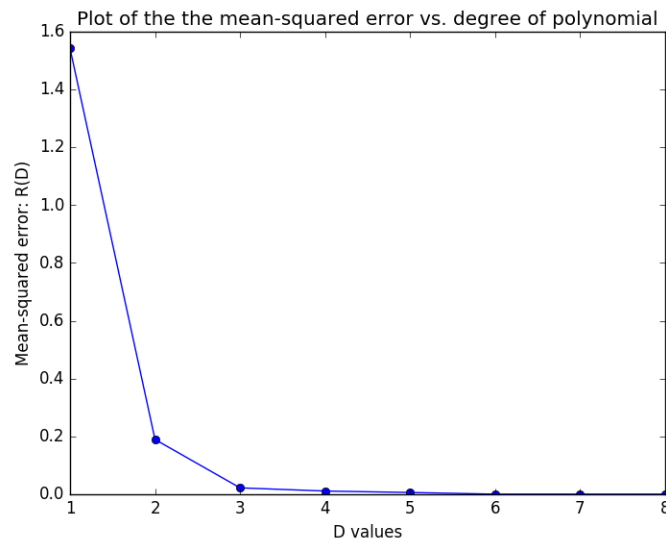


Figure 1: D vs. $R(D)$

(c) How does the MSE behave as a function of D and why? With the degree $n-1$ fit, we get (approximately) zero mean-squared error since the function fits exactly to every data point. What happens if you try to fit a polynomial of degree n ? Why? To fit a polynomial of degree n , we will

be solving $X\alpha = y$, where $X^{n \times n}$.

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

Using ordinary least-squares, we solve for $\alpha = (X^T X)^{-1} X^T y$.

(d) Figure 2 shows a plot of the degree $D \in 1, 2, \dots, n-1$ versus the mean-squared error $R(D)$ and \tilde{R} when using the data in y.dat, yfresh.dat, and t.dat. **Why do you think that this plot is qualitatively different from the plot in part (b)?** Even though we are fitting $D = n-1$ degree polynomial to the new yfresh.dat data, the model was trained on y.dat and will approximate yfresh.dat with greater error than the data it was trained on and cannot be a perfect estimator. Thus, the error appears to plateau for the same values of D with yfresh.dat or y.dat but at a higher error value when using yfresh.dat. **What does this tell you how the fitted degree D should be chosen?** Choose the minimal degree D after which the error doesn't change within some small ϵ bound.

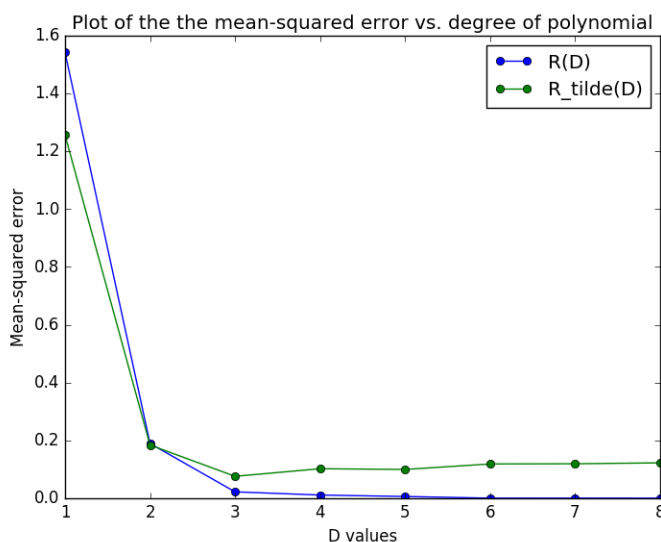


Figure 2: D vs. $R(D)$ and \tilde{R}

(e) Figure 3 shows a plot of the degree $D \in 2, \dots, 9$ versus the mean-squared error \tilde{R} and $F(D)$ when using the data in y.dat, yfresh.dat, and t.dat. **How are the minimizing arguments of the two functions related? Why is this an interesting observation?**

Problem 2.2.

(a) Prove that A is a convex function.

Proof. By definition,

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

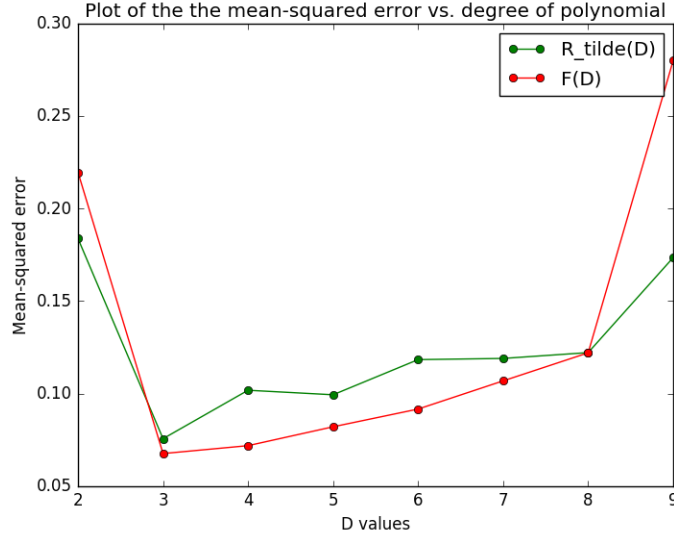


Figure 3: D vs. \tilde{R} and $F(D)$

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:

$$\begin{aligned} \frac{\partial}{\partial \eta} \frac{B'(\eta)}{B(\eta)} &= \frac{\partial}{\partial \eta} \left(B'(\eta) \frac{1}{B(\eta)} \right) = \frac{B''(\eta)}{B(\eta)} - \frac{(B'(\eta))^2}{B(\eta)^2} \\ &= \frac{\int_{\gamma} h(y) e^{\eta y} y^2 dy}{\int_{\gamma} h(y) e^{\eta y} dy} - (E_{p_{\eta}}[y])^2 = \frac{\int_{\gamma} h(y) e^{\eta y - A(\eta)} y^2 dy}{\int_{\gamma} h(y) e^{\eta y - A(\eta)} dy} - (E_{p_{\eta}}[y])^2 \\ &= E_{p_{\eta}}[y^2] - (E_{p_{\eta}}[y])^2 = \text{Var}_{p_{\eta}}[y] \succeq 0 \end{aligned}$$

Since $\text{Var}_{p_{\eta}}$ is positive definite, we have shown that $A(\eta)$ is convex. □

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

$$\begin{aligned} D(p_{\eta} || p_{\tilde{\eta}}) &= E_{\eta} \left(\log \left(\frac{h(y) e^{\eta y - A(\eta)}}{h(y) e^{\tilde{\eta} y - A(\tilde{\eta})}} \right) \right) \\ &= \int_y \log \left(e^{(\eta - \tilde{\eta})y - (A(\eta) - A(\tilde{\eta}))} p_{\eta}(y) \right) dy \\ &= \int_y ((\eta - \tilde{\eta})y - (A(\eta) - A(\tilde{\eta}))) h(y) e^{\eta y - A(\eta)} dy \end{aligned}$$

$$\begin{aligned}
&= (\eta - \tilde{n}) \int_y h(y) e^{\eta y - A(\eta)} y dy - (A(\eta) - A(\tilde{\eta})) \int_y h(y) e^{\eta y - A(\eta)} dy \\
&= (\eta - \tilde{n}) A' - A(\eta) + A(\tilde{\eta})
\end{aligned}$$

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:

$$\begin{aligned}
p_\eta(y) &= \eta^y (1 - \eta)^{1-y}, y \in 0, 1, n \in (0, 1) \\
&= e^{y \log(\eta) + (1-y) \log(1-\eta)} = e^{y \log(\frac{\eta}{1-\eta}) - \log(1+e^{\frac{\eta}{1-\eta}})}
\end{aligned}$$

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 .

$$\begin{aligned}
\nabla_\eta A^*(t) &= t - \frac{e^\eta}{1 + e^\eta} \\
0 &= t - \frac{e^\eta}{1 + e^\eta} \implies t = \frac{e^\eta}{1 + e^\eta} \\
\frac{1}{t} &= \frac{1 + e^\eta}{e^\eta} = \frac{1}{e^\eta} + 1 \implies \frac{1}{t} - 1 = \frac{1}{e^\eta} \\
e^\eta &= \frac{1}{\frac{1}{t} - 1} \implies \eta = \log\left(\frac{1}{\frac{1}{t} - 1}\right) \\
\eta &= -\log\left(\frac{1}{t} - 1\right)
\end{aligned}$$

Substituting this back into our equation, we get:

$$\begin{aligned}
A^*(t) &= -t \log\left(\frac{1}{t} - 1\right) - \log(1 + e^{-\log(\frac{1}{t} - 1)}) = -t \log\left(\frac{1}{t} - 1\right) + \log(1 - t) \\
&= t \log(t) - t \log(1 - t) + \log(1 - t) = t \log(t) + (1 - t) \log(1 - t)
\end{aligned}$$

(ii) Gaussian random variable:

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

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

$$\begin{aligned}
\nabla_\eta A^*(t) &= t - \eta \\
0 &= t - \eta \implies t = \eta
\end{aligned}$$

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 $A^*(t) = \sup_{\eta \in R} \{\eta t - e^\eta\}$.

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

$$0 = t - e^\eta \implies \log(t) = \eta$$

Substituting this back into our equation, we get:

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

(d) **Prove that conjugate dual is always a convex function**

Proof.

□

Problem 2.3.

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

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

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

$$\begin{aligned} \frac{\partial}{\partial \eta} l(\eta; y_1, \dots, y_n) &= \sum_{i=1}^n y_i - n \frac{\partial}{\partial \eta} A(\eta) \\ \frac{\partial}{\partial \eta} A(\eta) &= \frac{\sum_{i=1}^n y_i}{n} \implies \hat{\eta} = (A'^{-1})\left(\frac{\sum_{i=1}^n y_i}{n}\right) \end{aligned}$$

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

Poisson

$$\begin{aligned} \frac{\partial}{\partial \eta} A(\eta) &= e^\eta = \frac{\sum_{i=1}^n y_i}{n} \\ \hat{\eta} &= \log\left(\frac{\sum_{i=1}^n y_i}{n}\right) \end{aligned}$$

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

$$\begin{aligned} \frac{\partial}{\partial \eta} A(\eta) &= \frac{e^\eta}{1 + e^\eta} = \frac{\sum_{i=1}^n y_i}{n} \\ \hat{\eta} &= \log\left(\frac{\frac{\sum_{i=1}^n y_i}{n}}{1 - \frac{\sum_{i=1}^n y_i}{n}}\right) \end{aligned}$$

Gaussian

$$\begin{aligned} \frac{\partial}{\partial \eta} A(\eta) &= \eta = \frac{\sum_{i=1}^n y_i}{n} \\ \hat{\eta} &= \frac{\sum_{i=1}^n y_i}{n} \end{aligned}$$

(c)

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

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

Thus, we know that as $n \rightarrow \infty$, then $\bar{y} \rightarrow A'(\eta^*)$. 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:

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

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

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

(d) Assume A is strictly convex.

Problem 2.4.

(a) Based on the definition of MLE in GLM as well as stochastic gradient, we can redefine the function $L(\theta)$, the gradient $\Delta^t L(\theta)$, and the $\tilde{\theta}^{t+1}$ update step as:

$$\begin{aligned} L(\theta) &= -y_I x_I^T \theta^t + A(x_I^T \theta^t) \\ \Delta^t L(\theta) &= -y_I x_I^T + x_I^T A'(x_I^T \theta^t) \\ \tilde{\theta}^{t+1} &= \hat{\theta}^t - \gamma^t \Delta^t L(I) \end{aligned}$$

(b) Explicit updates for Poisson and Logistic cases:

Poisson

$$A(t) = e^t \implies \tilde{\theta}^{t+1} = \hat{\theta}^t - \gamma^t x_I^T e^{x_I^T \hat{\theta}^t}$$

Logistic

$$A(t) = \log(1 + e^t) \implies \tilde{\theta}^{t+1} = \hat{\theta}^t - \gamma^t x_I^T \left(\frac{e^{x_I^T \hat{\theta}^t}}{1 + e^{x_I^T \hat{\theta}^t}} \right)$$

(c) Figure ?? 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.

(d) Figure ?? 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 ?? and ?? are [] and suggest about the accuracy of the fits.

(e) (f)

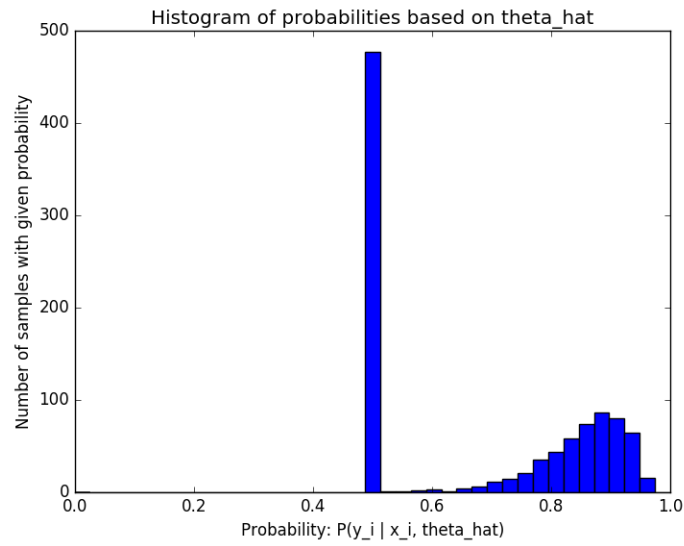


Figure 4: Histogram of probabilities using Xone.dat and yone.dat

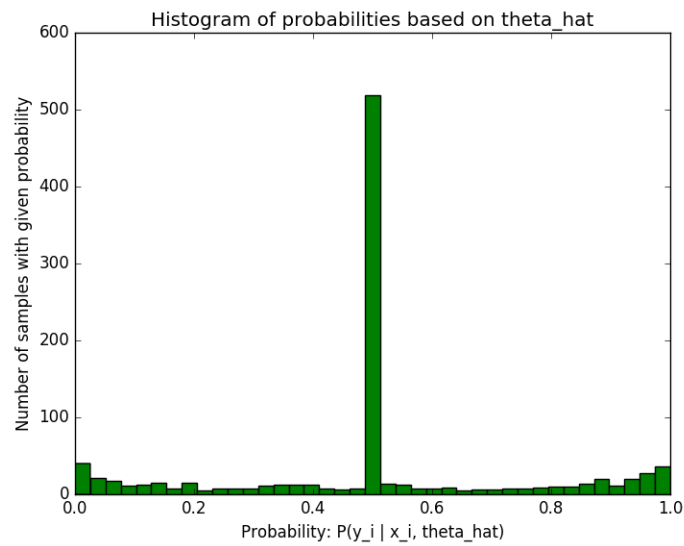


Figure 5: Histogram of probabilities using Xtwo.dat and ytwo.dat