CS229 Problem Set 1

anjay.h.friedman

July 2020

1 Introduction

This document represents my solutions to the 2017 CS229 course's first problem set. I appreciate that the materials for the course are available from Stanford's and other websites and hope that I am not infringing on any rights.

It took me 15 hours to complete and felt like an exercise in Deliberate Practice throughout.

2 Problems

2.1 1. [25 points] Logistic regression

(a) [10 points] Consider the average empirical loss (the risk) for logistic regression:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\log(1+e^{-y^{(i)}\theta^T x^{(i)}}) = -\frac{1}{m}\sum_{i=1}^{m}\log(h_\theta(y^{(i)}x^{(i)}))$$

where $y^{(i)} \in \{-1, 1\}$, $h_\theta(x) = g(\theta^T x)$ and $g(z) = 1/(1+e^{-z})$. Find the Hessian H of this function, and show that for any vector z, it holds true that

$$z^T H z \geq 0.$$

Hint: Be careful that the range for label values, $\{-1, 1\}$, is different than the range used in lecture notes, which is $\{0, 1\}$. Please read the supplementary notes if you are having trouble. You might want to start by showing the fact that $\sum_i \sum_j z_i x_i x_j z_j = (x^T z)^2 \geq 0$.
Remark: This is one of the standard ways of showing that the matrix H is positive semi-definite, written "$H \succeq 0$." This implies that J is convex, and has no local minima other than the global one.[1] If you have some other way of showing $H \succeq 0$, you're also welcome to use your method instead of the one above.

Solution

To start with, note that the Hessian H of a function f: $\mathbb{R}^{\aleph} \longmapsto \mathbb{R}$ is defined as the matrix of partial derivatives $\nabla_x^2 f(x)$. Thus $H_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$

The Hessian H of $J(\theta)$ is thus defined as the matrix where $H_{ij} = \frac{\partial^2 J(\theta)}{\partial\theta_i \partial\theta_j}$

Solving ... $\frac{\partial^2 J(\theta)}{\partial\theta_i\partial\theta_j} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}(-\frac{1}{m}\sum_{k=1}^{m} log(h_\theta(y^{(k)}x^{(k)})))$

$= -\frac{1}{m}\frac{\partial}{\partial\theta_j}(\sum_{k=1}^{m}\frac{\partial}{\partial\theta_i}log(h_\theta(y^{(k)}x^{(k)})))$ (1)

$= -\frac{1}{m}\frac{\partial}{\partial\theta_j}\sum_{k=1}^{m}\frac{1}{h_\theta(y^{(k)}x^{(k)})}\frac{\partial}{\partial\theta_i}(h_\theta(y^{(k)}x^{(k)}))$ (2)

$= -\frac{1}{m}\frac{\partial}{\partial\theta_j}\sum_{k=1}^{m}\frac{1}{g(y^{(k)}\theta^T x^{(k)})}\frac{\partial}{\partial\theta_i}(g(y^{(k)}\theta^T x^{(k)}))$ (3)

$= -\frac{1}{m}\frac{\partial}{\partial\theta_j}\sum_{k=1}^{m}\frac{1}{\cancel{g(y^{(k)}\theta^T x^{(k)})}}\cancel{g(y^{(k)}\theta^T x^{(k)})}(1-g(y^{(k)}\theta^T x^{(k)})\frac{\partial}{\partial\theta_i}(y^{(k)}\theta^T x^{(k)})$ (4)

$= -\frac{1}{m}\frac{\partial}{\partial\theta_j}\sum_{k=1}^{m}(1-g(y^{(k)}\theta^T x^{(k)}))y^{(k)}x_i^{(k)}$ (5)

$= \frac{1}{m}\sum_{k=1}^{m} y^{(k)}x_i^{(k)}\frac{\partial}{\partial\theta_j}(g(y^{(k)}\theta^T x^{(k)}))$ (6)

$= \frac{1}{m}\sum_{k=1}^{m} y^{(k)}x_i^{(k)}(g(y^{(k)}\theta^T x^{(k)}))(1-g(y^{(k)}\theta^T x^{(k)}))y^{(k)}x_j^{(k)}$ (7)

$= \frac{1}{m}\sum_{k=1}^{m}(y^{(k)})^2 x_i^{(k)}x_j^{(k)}g(y^{(k)}\theta^T x^{(k)})(1-g(y^{(k)}\theta^T x^{(k)}))$ (8)

$\implies \quad H_{ij} = \frac{1}{m}\sum_{k=1}^{m}(y^{(k)})^2 x_i^{(k)}x_j^{(k)}g(y^{(k)}\theta^T x^{(k)})(1-g(y^{(k)}\theta^T x^{(k)}))$ $\diamond$

(1) — pulling constants out and derivative of sum is sum of derivative (2) — chain rule for composite functions (3) — substitute g or the logistic function for h (4) — chain rule, g' is g(1-g) and cancel out (5) — derivative of $\theta^T x^{(k)}$ wrt θ_i is $x_i^{(k)}$ (6) — re-balancing and pulling constants out + $\partial\sum = \sum\partial$ (7) — derivative of logistic function chain rule + (5) (8) — re-balancing

To show that for any vector z, it holds true that $z^T Hz \geq 0$ or that H is *positive-semi definite*, we will first show that $\sum_i\sum_j z_i x_i x_j z_j = (x^T z)^2 \geq 0$

Proof:

$\sum_i\sum_j z_i x_i x_j z_j = \sum_i z_i x_i \sum_j x_j z_j = \sum_i z_i x_i (x^T z) = x^T z\sum_i z_i x_i = (x^T z)(x^T z)$

$= (x^T z)^2 \geq 0$ since $x^T z \in \mathbb{R}$ $\diamond$

Now that we have this, we will use it when showing that H is *positive semi-definite*

Claim: $z^T Hz \geq 0 \forall z \in \mathbb{R}^n$

Proof:

$z^T Hz = \sum_{i=1}^{n}\sum_{j=1}^{n} H_{ij}z_i z_j$ (1)

$= \sum_{i=1}^{n}\sum_{j=1}^{n} z_i z_j \frac{1}{m}\sum_{k=1}^{m}(y^{(k)})^2 x_i^{(k)}x_j^{(k)}g(y^{(k)}\theta^T x^{(k)})(1-g(y^{(k)}\theta^T x^{(k)}))$ (2)

$= \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m} z_i z_j (y^{(k)})^2 x_i^{(k)}x_j^{(k)}g(y^{(k)}\theta^T x^{(k)})(1-g(y^{(k)}\theta^T x^{(k)}))$ (3)

$= \frac{1}{m}\sum_{k=1}^{m}(y^{(k)})^2 g(y^{(k)}\theta^T x^{(k)})(1-g(y^{(k)}\theta^T x^{(k)}))\sum_{i=1}^{n}\sum_{j=1}^{n} z_i x_i^{(k)} z_j x_j^{(k)}$ (4)

$= \frac{1}{m}\sum_{k=1}^{m}(y^{(k)})^2 g(y^{(k)}\theta^T x^{(k)})(1-g(y^{(k)}\theta^T x^{(k)}))(x^{kT} z)^2$ (5)

≥ 0 (6)

$\implies \quad z^T Hz \geq 0 \forall z \in \mathbb{R}^n$ $\diamond$

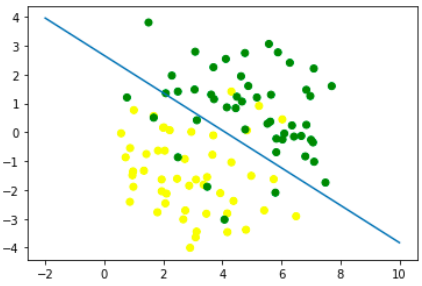
(1) — Can be shown by simply expanding (2) — Substituting in H (3) — Pulling out constants and simplifying (4) — Switching order of sums and pulling out some factors from inner sums (5) — Using the previous proof (6) — Follows since the sigmoid function, g, is greater than or equal to 0 and other numbers are squared

(b) [10 points] We have provided two data files:

- `http://cs229.stanford.edu/ps/ps1/logistic_x.txt`
- `http://cs229.stanford.edu/ps/ps1/logistic_y.txt`

These files contain the inputs ($x^{(i)} \in \mathbb{R}^2$) and outputs ($y^{(i)} \in \{-1, 1\}$), respectively for a binary classification problem, with one training example per row. Implement[2] Newton's method for optimizing $J(\theta)$, and apply it to fit a logistic regression model to the data. Initialize Newton's method with $\theta = \vec{0}$ (the vector of all zeros). What are the coefficients θ resulting from your fit? (Remember to include the intercept term.)

(c) [5 points] Plot the training data (your axes should be x_1 and x_2, corresponding to the two coordinates of the inputs, and you should use a different symbol for each point plotted to indicate whether that example had label 1 or -1). Also plot on the same figure the decision boundary fit by logistic regression. (This should be a straight line showing the boundary separating the region where $h_\theta(x) > 0.5$ from where $h_\theta(x) \leq 0.5$.)



Solution

Code for part (b) and (c) found here

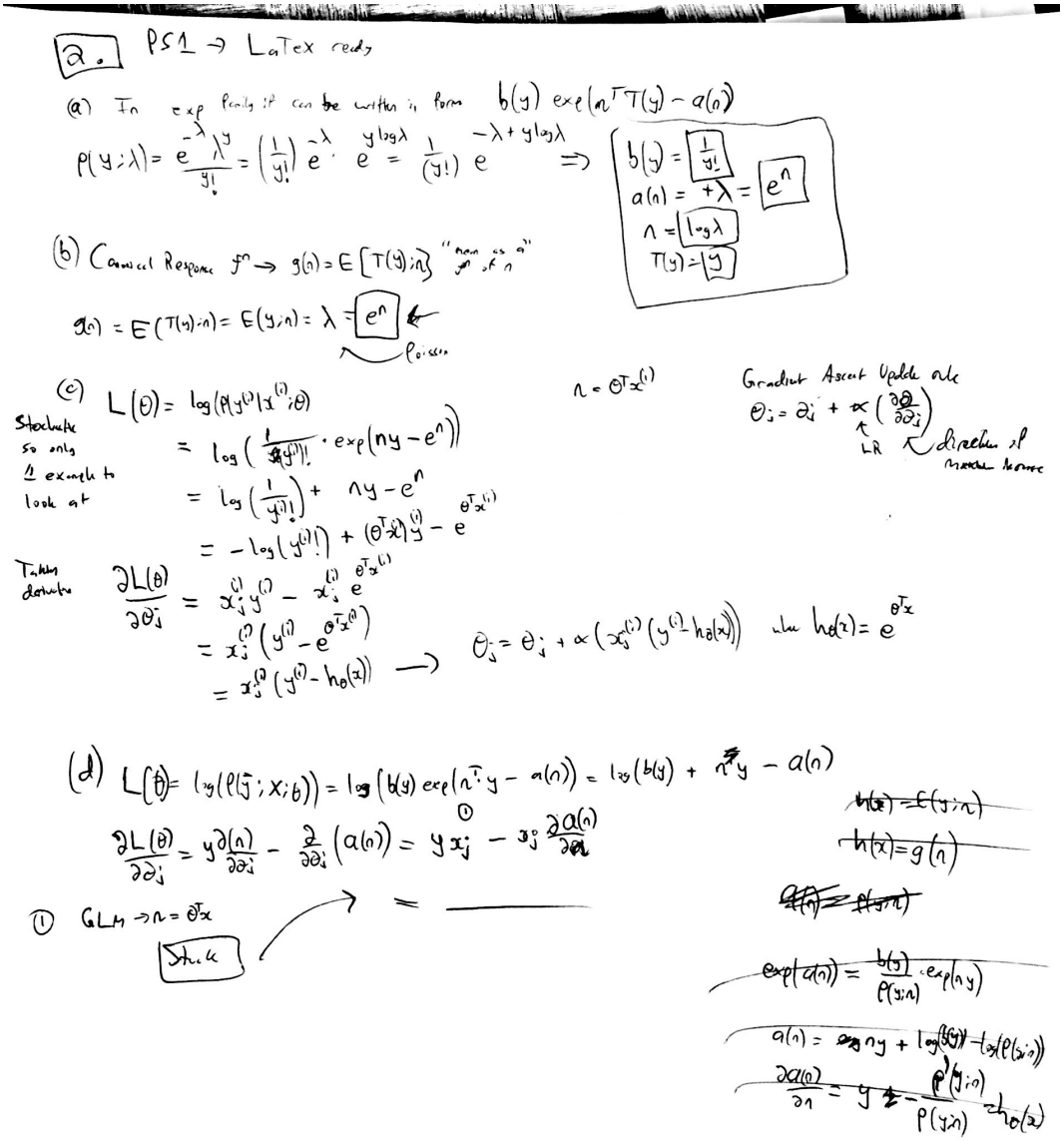
2.2 2. [15 points] Poisson regression and the exponential family

(a) [5 points] Consider the Poisson distribution parameterized by λ:

$$p(y; \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

Show that the Poisson distribution is in the exponential family, and clearly state what are $b(y)$, η, $T(y)$, and $a(\eta)$.

(b) [3 points] Consider performing regression using a GLM model with a Poisson response variable. What is the canonical response function for the family? (You may use the fact that a Poisson random variable with parameter λ has mean λ.)



2.3 3. [15 points] Gaussian discriminant analysis

Suppose we are given a dataset $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, m\}$ consisting of m independent examples, where $x^{(i)} \in \mathbb{R}^n$ are n-dimensional vectors, and $y^{(i)} \in \{-1, 1\}$. We will model the joint distribution of (x, y) according to:

$$\begin{aligned} p(y) &= \begin{cases} \phi & \text{if } y = 1 \\ 1 - \phi & \text{if } y = -1 \end{cases} \\ p(x|y = -1) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_{-1})^T \Sigma^{-1}(x - \mu_{-1})\right) \\ p(x|y = 1) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right) \end{aligned}$$

Here, the parameters of our model are ϕ, Σ, μ_{-1} and μ_1. (Note that while there're two different mean vectors μ_{-1} and μ_1, there's only one covariance matrix Σ.)

(a) [5 points] Suppose we have already fit ϕ, Σ, μ_{-1} and μ_1, and now want to make a prediction at some new query point x. Show that the posterior distribution of the label at x takes the form of a logistic function, and can be written

$$p(y \mid x; \phi, \Sigma, \mu_{-1}, \mu_1) = \frac{1}{1 + \exp(-y(\theta^T x + \theta_0))},$$

where $\theta \in \mathbb{R}^n$ and the bias term $\theta_0 \in \mathbb{R}$ are some appropriate functions of $\phi, \Sigma, \mu_{-1}, \mu_1$. (Note: the term θ_0 corresponds to introducing an extra coordinate $x_0^{(i)} = 1$, as we did in class.)

Solution

(b) [10 points] For this part of the problem only, you may assume n (the dimension of x) is 1, so that $\Sigma = [\sigma^2]$ is just a real number, and likewise the determinant of Σ is given by $|\Sigma| = \sigma^2$. Given the dataset, we claim that the maximum likelihood estimates of the parameters are given by

$$\begin{aligned}
\phi &= \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)} = 1\} \\
\mu_{-1} &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = -1\} x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = -1\}} \\
\mu_1 &= \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}} \\
\Sigma &= \frac{1}{m}\sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T
\end{aligned}$$

The log-likelihood of the data is

$$\begin{aligned}
\ell(\phi, \mu_{-1}, \mu_1, \Sigma) &= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \phi, \mu_{-1}, \mu_1, \Sigma) \\
&= \log \prod_{i=1}^{m} p(x^{(i)} | y^{(i)}; \mu_{-1}, \mu_1, \Sigma) p(y^{(i)}; \phi).
\end{aligned}$$

By maximizing ℓ with respect to the four parameters, prove that the maximum likelihood estimates of ϕ, μ_{-1}, μ_1, and Σ are indeed as given in the formulas above. (You may assume that there is at least one positive and one negative example, so that the denominators in the definitions of μ_{-1} and μ_1 above are non-zero.)

(c) **[3 extra credit points]** Without assuming that $n = 1$, show that the maximum likelihood estimates of ϕ, μ_{-1}, μ_1, and Σ are as given in the formulas in part (b). [Note: If you're fairly sure that you have the answer to this part right, you don't have to do part (b), since that's just a special case.]

Solution

2.4 4. [10 points] Linear invariance of optimization algorithms

Consider using an iterative optimization algorithm (such as Newton's method, or gradient descent) to minimize some continuously differentiable function $f(x)$. Suppose we initialize the algorithm at $x^{(0)} = \vec{0}$. When the algorithm is run, it will produce a value of $x \in \mathbb{R}^n$ for each iteration: $x^{(1)}, x^{(2)}, \ldots$.

Now, let some non-singular square matrix $A \in \mathbb{R}^{n \times n}$ be given, and define a new function $g(z) = f(Az)$. Consider using the same iterative optimization algorithm to optimize g (with initialization $z^{(0)} = \vec{0}$). If the values $z^{(1)}, z^{(2)}, \ldots$ produced by this method necessarily satisfy $z^{(i)} = A^{-1}x^{(i)}$ for all i, we say this optimization algorithm is **invariant to linear reparameterizations**.

For reference, Newton's Method is an algorithm for finding the minimum of a function based on its derivative.

Newton-Raphson Method $\theta := \theta - H^{-1}\nabla_\theta f(\theta)$ where H is the *Hessian*

(a) [7 points] Show that Newton's method (applied to find the minimum of a function) is invariant to linear reparameterizations. Note that since $z^{(0)} = \vec{0} = A^{-1}x^{(0)}$, it is sufficient to show that if Newton's method applied to $f(x)$ updates $x^{(i)}$ to $x^{(i+1)}$, then Newton's method applied to $g(z)$ will update $z^{(i)} = A^{-1}x^{(i)}$ to $z^{(i+1)} = A^{-1}x^{(i+1)}$.[3]

Outline of Solution:
1. Prove $\nabla_z g(z) = A^T \nabla_x f(x)$
2. Prove $H_z^{-1} = A^{-1} H_x^{-1} A^{-T}$
3. Substitute into $z^{(i+1)} = z^{(i)} - H_z^{-1}\nabla_z g(z)$ and show that it equals $A^{-1}x^{(i+1)}$ given that $z^{(i)} = A^{-1}x^{(i)}$

Solution

Claim: $\nabla_z g(z) = A^T \nabla_x f(x)$

Proof:

$\nabla_z g(z) = \frac{\partial(g(z))}{\partial z_i} = \frac{\partial}{\partial z_i}(f(Az))$
$= \frac{\partial}{\partial z_i} f(x)$ since $x = Az$
$= \sum_{k=1}^{n} \frac{\partial f(x)}{\partial x_k} \frac{\partial x_k}{\partial z_i}$ by multi-variable chain rule
$= \sum_{k=1}^{n} (\nabla_x f(x))_k A_{ki}$ since $x_k = (A_{k,}) * z$ and $\frac{\partial x_k}{\partial z_i} = A_{ki}$
$= < \nabla_x f(x), i^{th} col of A >= A_i^T \nabla_x f(x)$
$\implies \quad \nabla_z g(z) = A^T \nabla_x f(x) \quad \diamond$

Claim: $H_z^{-1} = A^{-1} H_x^{-1} A^{-T}$

Proof:

First we will solve for H_z using the previous result and then we will solve for the inverse.

$(H_z)_{ij} = \frac{\partial^2 g(z)}{\partial z_j \partial z_i}$ (definition of H)

$= \frac{\partial}{z_j}(\nabla_z g(z))_i = \frac{\partial}{\partial z_j}(A_i^T \nabla_x f(x)) = \frac{\partial}{\partial z_j}(\sum_{k=1}^n (\nabla_x f(x))_k A_{ki})$ (previous result)
$= \sum_{l=1}^n \frac{\partial}{\partial x_l}(\sum_{k=1}^n (\nabla_x f(x))_k A_{ki})(\frac{\partial x_l}{\partial z_j})$ (chain rule)
$= \sum_{l=1}^n A_{lj} \sum_{k=1}^n \frac{\partial}{\partial x_l}(\nabla_x f(x))_k A_{ki}$ (previous result and simplifying)
$= \sum_{l=1}^n \sum_{k=1}^n A_{lj} A_{ki} \frac{\partial^2 f(x)}{\partial x_l \partial x_k}$ (pulling constants out)
$= \sum_{l=1}^n \sum_{k=1}^n A_{lj} A_{ki} H_{kl}$ (definition of H)
$= \sum_{l=1}^n \sum_{k=1}^n (A_j^T)_l H_{kl} (A_i^T)_k = (A_j^T) H_x A_i = (A_i^T) H_x A_j$(H symmetric)
$\implies \quad H_z = A^T H_x A \implies \quad H_z^{-1} = A^{-1} H_x^{-1} A^{-T} \quad \diamond$

As the question states, to prove the answer by induction, all we have to show is that $z^{(i+1)} = A^{-1} x^{(i+1)}$ given that $z^{(i)} = A^{-1} x^{(i)}$ since it is obvious that $z^0 = A^{-1} x^0 = 0$

Formally, **Claim:** $z^{(i+1)} = A^{-1} x^{(i+1)}$
Proof:
$z^{(i+1)} = z^{(i)} - H_z^{-1} \nabla_z g(z)$ (update rule)
$= A^{-1} x^{(i)} - (A^{-1} H_x^{-1} A^{-T})(A^T \nabla_x f(x))$ (substitute previous results)
$= A^{-1} x^{(i)} - A^{-1} H_x^{-1} \nabla_x f(x)$ (simplify)
$= A^{-1}(x^{(i)} - H_x^{-1} \nabla_x f(x))$ (factor out A^{-1})
$\implies \quad = A^{-1}(x^{(i+1)}) \quad \diamond$

(b) [3 points] Is gradient descent invariant to linear reparameterizations? Justify your answer.

Solution
The Gradient Descent update rule is: $x^{(i+1)} = x^{(i)} - \alpha(\frac{\partial f(x)}{\partial x})$

Similar to part (a), it would be true if $z^{(i+1)} = z^{(i)} - \alpha \frac{\partial g(z)}{\partial z}$
$= A^{-1} x^{(i)} - \alpha \nabla_z g(z)$ (substitute in previous results)
$= A^{-1} x^{(i)} - \alpha A^T \nabla_x f(x)$ (substitute in previous results)
$= A^{-1}(x^{(i)} - \alpha A A^T \nabla_x f(x))$ (factoring out)
$\implies \quad = A^{-1}(x^{(i+1)})$ only if $AA^T = I \quad \diamond$
Thus the answer is no, generally.

2.5 5. [35 points] Regression for denoising quasar spectra

Introduction. In this problem, we will apply a supervised learning technique to estimate the light spectrum of *quasars*. Quasars are luminous distant galactic nuclei that are so bright, their light overwhelms that of stars in their galaxies. Understanding properties of the spectrum of light emitted by a quasar is useful for a number of tasks: first, a number of quasar properties can be estimated from the spectra, and second, properties of the regions of the universe through which the light passes can also be evaluated (for example, we can estimate the density of neutral and ionized particles in the universe, which helps cosmologists understand the evolution and fundamental laws governing its structure). The *light spectrum* is a curve that relates the light's intensity (formally, lumens per square meter), or *luminous flux*, to its wavelength. Figure 1 shows an example of a quasar light spectrum, where the wavelengths are measured in Angstroms (Å), where 1Å$= 10^{-10}$ meters.

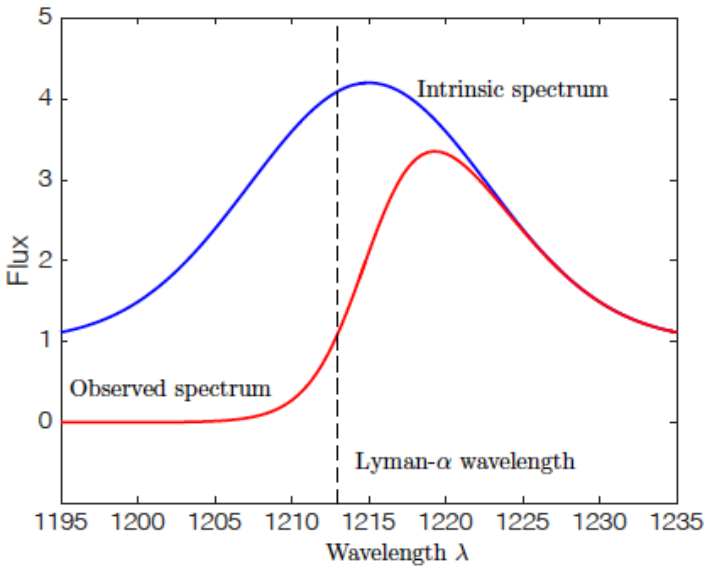


Figure 1: Light spectrum of a quasar. The blue line shows the intrinsic (i.e. original) flux spectrum emitted by the quasar. The red line denotes the observed spectrum here on Earth. To the left of the Lyman-α line, the observed flux is damped and the intrinsic (unabsorbed) flux continuum is not clearly recognizable (red line). To the right of the Lyman-α line, the observed flux approximates the intrinsic spectrum.

The Lyman-α wavelength is a wavelength beyond which intervening particles at most negligibly interfere with light emitted from the quasar. (Interference generally occurs when a photon is absorbed by a neutral hydrogen atom, which only occurs for certain wavelengths of light.) For wavelengths greater than this Lyman-α wavelength, the observed light spectrum f_{obs} can be modeled as a smooth spectrum f plus noise:

$$f_{\text{obs}}(\lambda) = f(\lambda) + \text{noise}(\lambda)$$

For wavelengths below the Lyman-α wavelength, a region of the spectrum known as the Lyman-α forest, intervening matter causes attenuation of the observed signal. As light emitted by the quasar travels through regions of the universe richer in neutral hydrogen, some of it is absorbed, which we model as

$$f_{\text{obs}}(\lambda) = \text{absorption}(\lambda) \cdot f(\lambda) + \text{noise}(\lambda)$$

Astrophysicists and cosmologists wish to understand the absorption function, which gives information about the Lyman-α forest, and hence the distribution of neutral hydrogen in otherwise unreachable regions of the universe. This gives clues toward the formation and evolution of the universe. Thus, it is our goal to estimate the spectrum f of an observed quasar.

Getting the data. We will be using data generated from the Hubble Space Telescope Faint Object Spectrograph (HST-FOS), Spectra of Active Galactic Nuclei and Quasars.[5] We have

(a) [10 points] Locally weighted linear regression

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

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

In class, we worked out what happens for the case where all the weights (the $w^{(i)}$'s) are the same. In this problem, we will generalize some of those ideas to the weighted setting.

i. [2 points] Show that $J(\theta)$ can also be written

$$J(\theta) = (X\theta - \vec{y})^T W (X\theta - \vec{y})$$

for an appropriate diagonal matrix W, and where X and $\vec{y}$ are as defined in class. State clearly what W is.

Solution

To solve this problem, I will assume there is some diagonal matrix W and then, expand the matrix representation of $J(\theta)$ to show that it is in fact equal to $J(\theta)$.

$J(\theta) = (X\theta - y)^T W (X\theta - y)$
$= \sum_{j=1}^{m}\sum_{k=1}^{m}(X\theta - y)_j W_{jk} (X\theta - y)_k$ (quadratic form expanded)
$= \sum_{i=1}^{m}(X\theta - y)_i^2 W_{ii}$ (W is diagonal so non-zero everywhere but j=k)
$= \sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2 W_{ii}$ (simplify)
Equal to $J(\theta)$ where $W_{ii} = \frac{w^{(i)}}{2}$ and $W_{jk} = 0$ if $j \neq k$ $\diamond$

ii. [4 points] If all the $w^{(i)}$'s equal 1, then we saw in class that the normal equation is

$$X^T X\theta = X^T \vec{y},$$

and that the value of θ that minimizes $J(\theta)$ is given by $(X^T X)^{-1} X^T \vec{y}$. By finding the derivative $\nabla_\theta J(\theta)$ and setting that to zero, generalize the normal equation to this weighted setting, and give the new value of θ that minimizes $J(\theta)$ in closed form as a function of X, W and $\vec{y}$.

Before solving this problem, I will first list some facts about matrix calculus and matrix transformations that will be used.

$\nabla_x(b^T x) = b$ $\quad$ $\nabla_x(x^T A x) = 2Ax$ $\quad$ $(a - b)T = (a^T - b^T)$
$a^T b = b^T a$ $\quad$ $(AB)^T = B^T A^T$

Solution

$J(\theta) = (X\theta - y)^T W (X\theta - y) = \theta^T X^T W X\theta - \theta^T X^T W y - y^T W X\theta + yWy$

$\nabla_\theta J(\theta) = \nabla_\theta(\theta^T X^T W X\theta) - \nabla_\theta((X^T W y)^T\theta) - \nabla_\theta((X^T W^T y)^T\theta) + \nabla_\theta(yWy)$
$0 = 2(X^T W X)\theta - X^T W y - X^T W^T y$
$2X^T W y = 2X^T W X\theta$
$\implies \quad \theta = (X^T W X)^{-1} X^T W y$ $\diamond$

iii. [4 points] Suppose we have a training set $\{(x^{(i)}, y^{(i)}); i = 1 \ldots, m\}$ of m independent examples, but in which the $y^{(i)}$'s were observed with differing variances. Specifically, suppose that

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

I.e., $y^{(i)}$ has mean $\theta^T x^{(i)}$ and variance $(\sigma^{(i)})^2$ (where the $\sigma^{(i)}$'s are fixed, known, constants). Show that finding the maximum likelihood estimate of θ reduces to solving a weighted linear regression problem. State clearly what the $w^{(i)}$'s are in terms of the $\sigma^{(i)}$'s.

Solution

I will solve this problem by showing that maximizing the log-likelihood probabilities is the same as minimizing a function that is the same as the LMS weighted cost function

Proof:

Maximizing $l(\theta) = \prod_{i=1}^{m} P(y^{(i)}|x^{(i)}; \theta)$ is the same as maximizing

$L(\theta) = log(\prod_{i=1}^{m} P(y^{(i)}|x^{(i)}; \theta)) = \sum_{i=1}^{m} log(P(y^{(i)}|x^{(i)}; \theta))$

$= \sum_{i=1}^{m} log(\frac{1}{\sqrt{2\pi}\sigma^{(i)}} exp(-\frac{(y_{(i)} - \theta^T x^{(i)})^2}{2\sigma_{(i)}}^2))$

$= \sum_{i=1}^{m} -log(\sqrt{2\pi}\sigma^{(i)}) - \frac{(y_{(i)} - \theta^T x^{(i)})^2}{2\sigma_{(i)}}^2$

Max $C - \sum_{i=1}^{m} \frac{(y_{(i)} - \theta^T x^{(i)})^2}{2\sigma_{(i)}}^2$ = Min $\sum_{i=1}^{m} \frac{(y_{(i)} - \theta^T x^{(i)})^2}{2\sigma_{(i)}}^2$

$\implies$ $= J(\theta)$ where $w^{(i)} = \frac{1}{(\sigma^{(i)})^2}$ $\diamond$

Rest of question and code solution found here