Instituto Tecnológico de Costa Rica
Escuela de Ingeniería Electrónica
Profesor: Dr. Pablo Alvarado Moya
EL-5852 Introducción al Reconocimiento de Patrones
I Semestre 2020

Entrega: **17 de marzo, 2020**

Tarea 3

Regresión para eliminar ruido de espectros de cuásares

por Andrew Ng

I Introducción

En este problema usted aplicará una técnica de aprendizaje supervisado en la estimación de el espectro de luz de cuásares. Cuásares son núcleos galácticos distantes, tan brillantes, que su luz oprime la de las estrellas en sus galaxias. Las propiedades del espectro de luz emitida por los cuásares permite: primero, estimar propiedades de los cuásares, y segundo, estimar propiedades de las regiones del universo que su luz atraviesa. Por ejemplo, podemos estimar la densidad de partículas neutrales y ionisadas en el universo, lo que permite a los cosmólogos comprender la evolución y las leyes fundamentales que gobiernan su estructura.

El espectro luminoso es una curva que relaciona el flujo lumínico (en lumens por metro cuadrado) con la longitud de onda. La figura 1 muestra un ejemplo del espectro luminoso de un cuásar, donde las longitudes de onda están medidas en angstroms (Å), con $1\,\text{Å} = 10^{-10}\,\text{m}$.

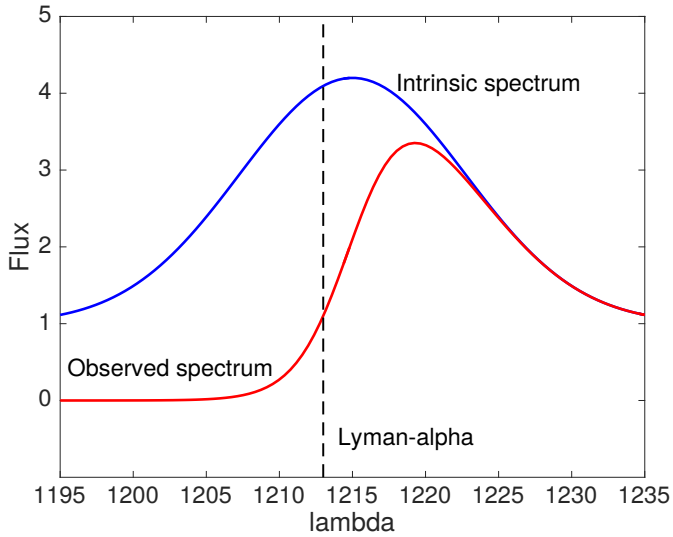

Figura 1: Espectro luminoso de un cuásar. La línea azul muestra el flujo espectral intrínseco (esto es, el original) emitido por el cuásar. La línea roja denota el espectro observado desde la Tierra. A la izquierda de la línea Lyman-α, el flujo observado está atenuado y el flujo intrínseco (no absorvido) no es reconocible claramente (línea roja). A la derecha de la línea Lyman-α el flujo observado aproxima al espectro intrínseco.

La longitud de onda Lyman-α es una longitud de onda sobre la cual partículas tienen interferencia despreciable con la luz emitida por el cuásar. La interferencia generalmente ocurre cuando un fotón es absorbido por un átomo neutral de hidrógeno, lo que ocurre solo para longitudes de onda

particulares. Para longitudes de onda superiores a la de Lyman-α, el espectro de luz observado f_{obs} puede ser modelado como un espectro suave f más ruido:

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

Para longitudes de onda bajo la de Lyman-α, en una región del espectro conocida como el bosque de Lyman-α, materia presente causa atenuación en la señal observada. Cuando la luz emitida por el cuásar viaja a través de regiones del universo más ricas en hidrógeno neutral, alguna de ellas es absorbida, lo que se modela como

$$f_{obs}(\lambda) = \text{absorción}(\lambda) \cdot f(\lambda) + \text{ruido}(\lambda)$$

En astrofísica y cosmología se desea comprender la función de absorción, que da información sobre el bosque de Layman-α, e indirectamente con ella conocer la distribución de hidrógeno neutral en zonas alejadas del universo. Nuestro objetivo es estimar el espectro f de un cuásar observado.

II Datos

Vamos a utilizar datos generados por el espectrógrafo de objetos tenues del Telescopio Espacial Hubble (HST-FOS, *Hubble Space Telescope Faint Object Spectrograph*), espectros de núcleos galácticos activos y cuásares[1]. En el tecDigital, en la carpeta de la Tarea 3, se han colocado dos archivos de datos `quasar_train.csv` y `quasar_test.csv`.

Cada archivo contiene una única fila de encabezado con 450 números correspondientes a longitudes de onda enteras en el intervalo $[1150, 1600]$ Å. El resto de las líneas contiene medidas relativas del flujo lumínico para cada longitud de onda. Específicamente, `quasar_train.csv` contiene 200 ejemplos y `quasar_test.csv` contiene 50 ejemplos. Usted puede utilizar `load_quasar_data.m` para cargar datos en GNU/Octave

III Tareas

1. Considere un problema de regresión lineal en el que queremos "ponderar" de forma distinta cada ejemplo de entrenamiento, tal y como vimos en clase con el método de regresión ponderada localmente. Específicamente queremos minimizar

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

   1.1. Demuestre que para el caso general $J(\underline{\boldsymbol{\theta}})$ se puede reescribir como

$$J(\underline{\boldsymbol{\theta}}) = \frac{1}{2}(\mathbf{X}\underline{\boldsymbol{\theta}} - \underline{\mathbf{y}})^T \mathbf{W} (\mathbf{X}\underline{\boldsymbol{\theta}} - \underline{\mathbf{y}})$$

   para una matriz diagonal $\mathbf{W}$ apropiada donde $\mathbf{X}$ es la matriz de diseño e $\underline{\mathbf{y}}$ el vector de salidas, tal y como lo definimos en clase.

[1] https://hea-www.harvard.edu/FOSAGN/

Al término $r_i = \underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}}^{(i)} - y^{(i)}$ *se le denomina* residuo. *El vector de residuos*

$$\underline{\mathbf{r}} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix} = \begin{bmatrix} \underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}}^{(1)} - y^{(1)} \\ \underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}}^{(2)} - y^{(2)} \\ \vdots \\ \underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}}^{(m)} - y^{(m)} \end{bmatrix} = \begin{bmatrix} \underline{\mathbf{x}}^{(1)T} \underline{\boldsymbol{\theta}} - y^{(1)} \\ \underline{\mathbf{x}}^{(2)T} \underline{\boldsymbol{\theta}} - y^{(2)} \\ \vdots \\ \underline{\mathbf{x}}^{(m)T} \underline{\boldsymbol{\theta}} - y^{(m)} \end{bmatrix} = \begin{bmatrix} \underline{\mathbf{x}}^{(1)T} \\ \underline{\mathbf{x}}^{(2)T} \\ \vdots \\ \underline{\mathbf{x}}^{(m)T} \end{bmatrix} \underline{\boldsymbol{\theta}} - \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} = \mathbf{X}\underline{\boldsymbol{\theta}} - \underline{\mathbf{y}}$$

Obsérvese que la función objetivo se puede reexpresar como

$$J(\underline{\boldsymbol{\theta}}) = \frac{1}{2} \sum_{i=1}^{m} w^{(i)} r_i^2 = \frac{1}{2} \sum_{i=1}^{m} r_i w^{(i)} r_i = \frac{1}{2} \underline{\mathbf{r}}^T \operatorname{diag}(\underline{\mathbf{w}}) \underline{\mathbf{r}}$$

lo que es idéntico a

$$J(\underline{\boldsymbol{\theta}}) = \frac{1}{2} (\mathbf{X}\underline{\boldsymbol{\theta}} - \underline{\mathbf{y}})^T \mathbf{W} (\mathbf{X}\underline{\boldsymbol{\theta}} - \underline{\mathbf{y}})$$

1.2. Si todos los pesos $w^{(i)}$ son iguales a 1, entonces en clase vimos que la ecuación normal es símplemente

$$\mathbf{X}^T \mathbf{X} \underline{\boldsymbol{\theta}} = \mathbf{X}^T \underline{\mathbf{y}}$$

y el valor de $\underline{\boldsymbol{\theta}}$ que minimiza $J(\underline{\boldsymbol{\theta}})$ está dado por $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underline{\mathbf{y}}$.
Encuentre el gradiente $\nabla_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}})$ e iguálelo a cero para encontrar una versión generalizada de las ecuaciones normales en este contexto con ponderación. Encuentre una forma cerrada del valor de $\underline{\boldsymbol{\theta}}$ que minimiza a $J(\underline{\boldsymbol{\theta}})$, en función de $\mathbf{X}$, $\mathbf{W}$ e $\underline{\mathbf{y}}$.
Podemos expander la función de error:

$$\begin{aligned} J(\underline{\boldsymbol{\theta}}) &= \frac{1}{2} (\underline{\boldsymbol{\theta}}^T \mathbf{X}^T - \underline{\mathbf{y}}^T) \mathbf{W} (\mathbf{X}\underline{\boldsymbol{\theta}} - \underline{\mathbf{y}}) \\ &= \frac{1}{2} (\underline{\boldsymbol{\theta}}^T \mathbf{X}^T - \underline{\mathbf{y}}^T) (\mathbf{W}\mathbf{X}\underline{\boldsymbol{\theta}} - \mathbf{W}\underline{\mathbf{y}}) \\ &= \frac{1}{2} \left(\underline{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{W} \mathbf{X} \underline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{W} \underline{\mathbf{y}} - \underline{\mathbf{y}}^T \mathbf{W} \mathbf{X} \underline{\boldsymbol{\theta}} + \underline{\mathbf{y}}^T \mathbf{W} \underline{\mathbf{y}} \right) \end{aligned}$$

que es un escalar, por lo que podemos hacer:

$$\nabla_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}}) = \nabla_{\underline{\boldsymbol{\theta}}} \operatorname{tr} \left\{ \frac{1}{2} \left(\underline{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{W} \mathbf{X} \underline{\boldsymbol{\theta}} - \underline{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{W} \underline{\mathbf{y}} - \underline{\mathbf{y}}^T \mathbf{W} \mathbf{X} \underline{\boldsymbol{\theta}} + \underline{\mathbf{y}}^T \mathbf{W} \underline{\mathbf{y}} \right) \right\}$$

y como el último término no depende de $\underline{\boldsymbol{\theta}}$

$$\nabla_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}}) = \frac{1}{2} \nabla_{\underline{\boldsymbol{\theta}}} \operatorname{tr} \left\{ \underline{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{W} \mathbf{X} \underline{\boldsymbol{\theta}} \right\} - \frac{1}{2} \nabla_{\underline{\boldsymbol{\theta}}} \operatorname{tr} \left\{ \underline{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{W} \underline{\mathbf{y}} \right\} - \frac{1}{2} \nabla_{\underline{\boldsymbol{\theta}}} \operatorname{tr} \left\{ \underline{\mathbf{y}}^T \mathbf{W} \mathbf{X} \underline{\boldsymbol{\theta}} \right\}$$

Para el primer término, y tomando los resultados de la primera tarea, sabemos que

$$\begin{aligned} \frac{1}{2} \nabla_{\underline{\boldsymbol{\theta}}} \operatorname{tr} \left\{ \underline{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{W} \mathbf{X} \underline{\boldsymbol{\theta}} \right\} &= \mathbf{X}^T \mathbf{W} \mathbf{X} \underline{\boldsymbol{\theta}} \\ \frac{1}{2} \nabla_{\underline{\boldsymbol{\theta}}} \operatorname{tr} \left\{ \underline{\boldsymbol{\theta}}^T \mathbf{X}^T \mathbf{W} \underline{\mathbf{y}} \right\} &= \frac{1}{2} \mathbf{X}^T \mathbf{W} \underline{\mathbf{y}} \\ \frac{1}{2} \nabla_{\underline{\boldsymbol{\theta}}} \operatorname{tr} \left\{ \underline{\mathbf{y}}^T \mathbf{W} \mathbf{X} \underline{\boldsymbol{\theta}} \right\} &= \frac{1}{2} \mathbf{X}^T \mathbf{W}^T \underline{\mathbf{y}} = \frac{1}{2} \mathbf{X}^T \mathbf{W} \underline{\mathbf{y}} \end{aligned}$$

donde se ha usado el hecho de que $\mathbf{W}$ *es diagonal, y por tanto* $\mathbf{W} = \mathbf{W}^T$*. Por lo tanto*

$$\nabla_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}}) = \mathbf{X}^T\mathbf{W}\mathbf{X}\underline{\boldsymbol{\theta}} - \mathbf{X}^T\mathbf{W}\underline{\mathbf{y}}$$

que debemos igualar a cero para encontrar el mínimo:

$$\begin{aligned}\mathbf{X}^T\mathbf{W}\mathbf{X}\underline{\boldsymbol{\theta}} - \mathbf{X}^T\mathbf{W}\underline{\mathbf{y}} &= 0\\ \mathbf{X}^T\mathbf{W}\mathbf{X}\underline{\boldsymbol{\theta}} &= \mathbf{X}^T\mathbf{W}\underline{\mathbf{y}}\\ \underline{\boldsymbol{\theta}} &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\underline{\mathbf{y}}\end{aligned}$$

1.3. Suponga que tenemos un conjunto de entrenamiento $\{(\underline{\mathbf{x}}^{(i)}, y^{(i)}); i = 1 \dots m\}$ de m ejemplos independientes, pero en el que los $y^{(i)}$ se observaron con varianzas distintas. Específicamente, suponga que

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

o en otras palabras, $y^{(i)}$ tiene media $\underline{\boldsymbol{\theta}}^T\underline{\mathbf{x}}^{(i)}$ y varianza $(\sigma^{(i)})^2$, donde las $\sigma^{(i)}$ son constantes fijas, conocidas. Demuestre que encontrar el estimado de máxima verosimilitud de $\underline{\boldsymbol{\theta}}$ se reduce a resolver un problema de regresión lineal ponderada. Establezca claramente que los $w^{(i)}$ se calculan en términos de $\sigma^{(i)}$.

La verosimilitud la calculamos con

$$L(\underline{\boldsymbol{\theta}}) = \prod_{i=1}^{m} p(y^{(i)} \mid \underline{\mathbf{x}}^{(i)}; \underline{\boldsymbol{\theta}})$$

y la verosimilitud logarítmica

$$\begin{aligned}\ell(\underline{\boldsymbol{\theta}}) = \ln L(\underline{\boldsymbol{\theta}}) &= \ln \prod_{i=1}^{m} p(y^{(i)} \mid \underline{\mathbf{x}}^{(i)}; \underline{\boldsymbol{\theta}})\\ &= \sum_{i=1}^{m} \ln p(y^{(i)} \mid \underline{\mathbf{x}}^{(i)}; \underline{\boldsymbol{\theta}})\\ &= \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi}\sigma^{(i)}} \exp\left(-\frac{(y^{(i)} - \underline{\boldsymbol{\theta}}^T\underline{\mathbf{x}}^{(i)})^2}{2(\sigma^{(i)})^2}\right)\\ &= \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi}\sigma^{(i)}} + \sum_{i=1}^{m} \left(-\frac{(y^{(i)} - \underline{\boldsymbol{\theta}}^T\underline{\mathbf{x}}^{(i)})^2}{2(\sigma^{(i)})^2}\right)\\ &= \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi}\sigma^{(i)}} - \frac{1}{2}\sum_{i=1}^{m} \frac{1}{(\sigma^{(i)})^2}(y^{(i)} - \underline{\boldsymbol{\theta}}^T\underline{\mathbf{x}}^{(i)})^2\end{aligned}$$

El primer término es una constante, y por tanto para maximizar la verosimilitud $\ell(\underline{\boldsymbol{\theta}})$ *se requiere minimizar el segundo término, que es idéntico a* $J(\underline{\boldsymbol{\theta}})$ *haciendo*

$$w^{(i)} = \frac{1}{(\sigma^{(i)})^2}$$

2. Visualización de los datos

   Nota: Observe que usted tiene a disposición el script `load_quasar_data.m` para cargar los datos en:

   - `lambdas`: las longitudes de onda,
   - `train_qso`: los datos de entrenamiento y
   - `test_qso`: con los datos de prueba

   2.1. Use las ecuaciones normales para implementar la regresión lineal (*no* ponderada) $y = \underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}}$ en el *primer* ejemplo de entrenamiento (esto es, la primera fila de `train_qso`). En una figura, grafique tanto los datos crudos como la línea recta resultante del ajuste. Indique el $\underline{\boldsymbol{\theta}}$ óptimo resultante de la regresión lineal.

   2.2. Implemente la regresión lineal ponderada localmente en el *primer* ejemplo de entrenamiento. Use las ecuaciones normales que usted derivó en el punto 1.2. En una figura aparte, grafique tanto los datos crudos como la curva suave resultante de su regresión. Cuando evalúe $h(\cdot)$ en un punto $\underline{\mathbf{x}}$ use los pesos

   $$w^{(i)} = \exp\left(-\frac{\|\underline{\mathbf{x}} - \underline{\mathbf{x}}^{(i)}\|^2}{2\tau^2}\right)$$

   con el parámetro de ancho de banda $\tau = 5$.

   2.3. Repita el punto 2.2 cuatro veces más con $\tau = 1, 10, 100$ y 1000. Grafique las curvas resultantes en una misma figura. Indique en una frase corta qué ocurre a la curva de regresión conforme τ crece.

IV Entregables

1. Archivo PDF con las demostraciones matemáticas de la primera parte, y las gráficas generadas para la segunda parte.
2. Archivos de GNU/Octave que usted realice para resolver los puntos anteriores
3. Archivo README con instrucciones de cómo ejecutar su código

V Notas

- Esta tarea es preferible resolverla en parejas. Envíe por favor lo antes posible a la cafetería del tecDigital, al hilo de grupos de la tarea 3 creado por el profesor, si usted va a trabajar solo o el nombre de su pareja de trabajo. ¡De otro modo no podrán subir sus tareas¡
- No es una tarea compleja, pero requiere tiempo para comprender/derivar la matemática y hacer el programa, que por lo general sale en relativamente pocas líneas.