Máster en Data Science. URJC
Sede Madrid

Inteligencia y Analítica de Negocios - GRUPO SEDE MADRID - INTENSIVO - A

Carlos Sánchez Vega

2017/2018

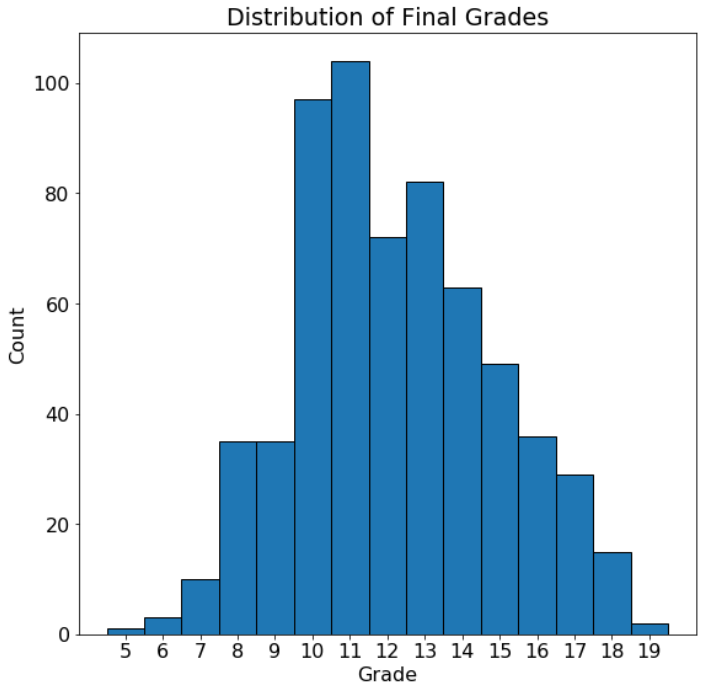
Índice

0.1 ¿Para qué variables hay que definir, en este caso, una distribución "a priori"? ¿Qué distribuciones se han escogido para esta implementación? ¿Por qué?

En el caso del ejemplo del tutorial, se hizo un analisis exploratorio del conjunto de datos (EDA). Después de estudiar la correlación de las variables (mediante los coeficientes de correlación), se llevó a cabo una reducción de la dimensionalidad y se seleccionaron 6 variables independientes (predictores). Esas seis variables son las que están más correlacionadas con las notas :

Grade ~ failures + higher_edu + mother_edu + studytime + father_edu + absences

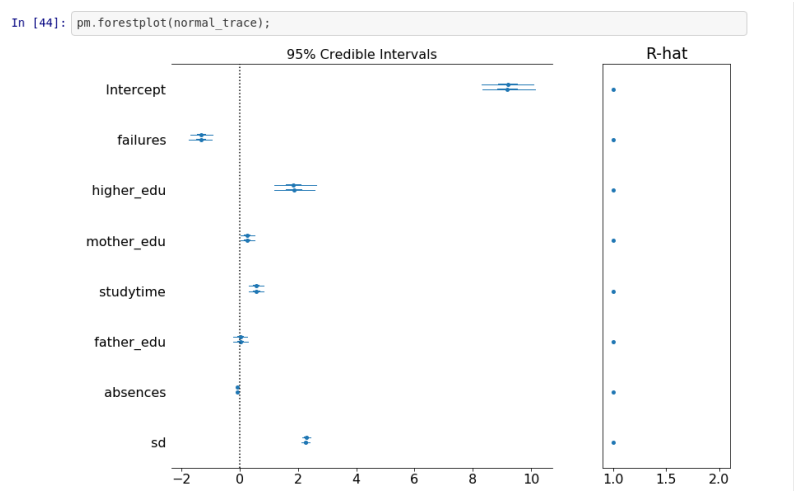
Es decir, las variables para las que hay que definir una distribución a priori son los predictores anteriormente mostrados: failures + higher_edu + mother_edu + studytime + father_edu + absences. Generalmente, el prior refleja la información previa que se posea respecto una distribución. Tal y como se mostró en el análisis EDA, la distribución para la media era la distribución normal:



La distribución a posteriori es de la misma familia que la distribución que la distribución a priori (si las distribución a posteriori no perteneciese a la misma familia que la distribución a priori, sería muy complejo actualizar el conocimiento sobre la distribución).

0.2 Escribe el significado del intervalo de confianza de la solución frecuentista del problema (un párrafo que indique cómo reportarías en un documento el resultado obtenido para dicho intervalo). A continuación, escribe el significado del p-valor obtenido para los dos coeficientes del modelo de regresión. A continuación, indica cómo reportarías el intervalo de confianza obtenido mediante la solución bayesiana aquí planteada.

Inicialmente, queremos saber el valor promedio de nuestras variables independientes. Como no disponemos de todos los datos de las notas de Portugal, sino un conjunto de datos de muestra, se usa el intervalo de confianza. El intervalo de confianza nos dará un rango de valores para los cuales el 95% de veces que se repita el análisis, con muestras de la población elegidas al azar, encontraremos los valores promedios de nuestras 6 variables priors anteriormente mostradas. Es decir, el intervalo de confianza nos va a permitir estimar entre qué valores está el valor inaccesible real de la población de las notas de los estudiantes, a partir de los datos que disponemos con nuestra muestra (con una probabilidad de equivocarnos un 5%)



En este gráfico se pueden ver lo valores más probables de los parámetros (el punto) junto con el 95% del intervalo de credibilidad para cada parámetro. Como se puede ver "higher_edu" es la que tiene mayor certeza (menor desviación) en comparación con otras variables.

El p-valor es una medida de la fuerza de la evidencia en sus datos en contra de Ho (hipótesis alternativa). Por lo general, cuanto más pequeño sea el p-valor, más fuerte será la evidencia de la muestra para rechazar Ho. Para los bayesianos, la probabilidad es un grado de creencia, y el enfoque sería el siguiente: se considera la curva que representa la función de densidad que se obtiene a posteriori, y si el área bajo dicha curva entre los valores X e Y es igual a 95%, entonces se puede hablar de que el verdadero valor esté entre X e Y con una probabilidad

del 95%. Es decir,el estimador de las notas a posteriori , teniendo la distribución normal de nuestra muestra, se va a encontrar entre dos valores X e Y el 95% de las veces.

0.3 ¿Qué algoritmo de muestreo estamos usando para calcular la distribución a posteriori? ¿Cómo se resuelve el problema de que no podamos conocer de forma analítica el factor en el denominador de la fórmula del teorema de Bayes?

Se usa el algoritmo de las "Cadenas de Markov Monte Carlo" para generar muestras, de la distribución a posteriori, con intención de aproximarla. Más concretamente, el algoritmo se podría dividir en: por un lado, "Monte Carlo" es la técnica por la que se generan muestras aleatorias y, por otro lado, las cadenas de "Markov" se apoyan en el concepto de que la siguiente muestra dependa del valor de la muestra previa. De este modo, a medida que se tengan más muestras, la aproximación de la distribución a posteriori convergerá a la verdadera distribución de los parámetros del modelo.

En lo que concierne al problema del donominador en la función de Bayes, si recordamos la función es la siguiente:

$$P(\beta|y,X) = \frac{P(y|\beta,X) * P(\beta|X)}{P(y|X)}$$

Siendo el multiplicando, la probabilidad de la variable dependiente (notas), el multiplicador la probabilidad de las variables independientes (los seis predictores) y denominador es la probabilidad de producción de esos datos. Como el denominador no depende de la variable dependiente (notas) y los valores de las variables independientes son datos, podemos decir que el denominador es constante y por tanto,se suele obviar.

0.4 ¿Qué significa el HPD que se representa gráficamente en las figuras para las distribuciones a posteriori de los coeficientes y de la desviación típica del modelo? Explica qué otros intervalos alternativos o estimadores se podrían utilizar para reportar el resultado (e.g. MAP). Indica también qué ventajas tiene el HPD frente a otras alternativas.

HPD (Highest Posterior Density) es un intervalo de credibilidad. Es decir, es el equivalente Bayesiano al intervalo de confianza en la hipótesis frecuentista (aunque con diferentes interpretaciones). Cualquier punto en el intervalo de credibilidad tiene mayor densidad que otro fuera, por lo que podríamos decir que el intervalo de valores de HPD es la colección de valores más probables de los parámetros a posteriori (es decir, son los valores más probables de los parámetros de nuestras distribuciones a posteriori). Por tanto, el HPD nos sirve para identificar el nivel de confianza en los parámetros del modelo. Si nos centramos en el conjunto de datos de las notas de Pôrtugal, podemos poner como ejemplo la siguiente la siguiente tabla:

	mean	sd	mc_error	hpd_2.5	hpd_97.5	n_eff	Rhat
Intercept	9.197350	0.466233	0.009974	8.339046	10.129492	2505.0	0.999956
failures	-1.316566	0.200428	0.003174	-1.701078	-0.917525	3879.0	0.999768
higher_edu	1.854784	0.361739	0.006817	1.202987	2.607551	3202.0	1.000270
mother_edu	0.263578	0.125141	0.002255	0.019157	0.509765	2942.0	0.999802
studytime	0.577236	0.127546	0.002195	0.326653	0.827031	3301.0	0.999762
father_edu	0.030319	0.128503	0.002093	-0.227607	0.272364	3020.0	0.999836
absences	-0.067602	0.023550	0.000372	-0.114305	-0.022007	3512.0	0.999754
sd	2.281731	0.076456	0.001325	2.141054	2.438024	3405.0	1.001020

La columna “father_edu” tiene un 95% de HPD que va de -0.22 a 0.27, unido a que la desviación estándar es muy grande, indica que no estamos muy seguros de los valores de las distribución a posteriori de dicho parámetro.

Por otro lado, MAP (Maximum a posteriori), es una estimación que sirve para tomar el valor que es más probable dado un conocimiento a priori.

La ventaja de HPD, frente a otras medidas, es que se calculan probabilidades para las funciones de densidad en lugar de apromimaciones en un punto concreto. Por otro lado, HPD es un método que sirve para incorporar información (tiene en cuenta información previa).