Departamento de Ingeniería Matemática
MA3402-1 Estadística
4 de diciembre 2019

Tarea final: Test de hipótesis, inferencia bayesiana y modelos lineales

Profesor: Felipe Tobar
Auxiliares: Diego Marchant, Francisco Vásquez
Fecha de entrega: 4 de enero 2020

Contexto: Esta tarea contiene partes teóricas y prácticas, donde el objetivo es evaluar su conocimiento sobre test de hipótesis, inferencia bayesiana y regresión lineal. Por esta razón, es fundamental que sus respuestas estén presentadas de forma clara y evidencien su entendimiento de los conceptos evaluados y su análisis científico.
Indicaciones: Esta tarea debe ser realizada en grupos de 1 o 2 alumnos y si bien puede ser producida en cualquier formato (LaTeX recomendado) debe ser entregada en formato `.pdf`. Esta vez no hay un número máximo de páginas permitido, sin embargo, la corrección tomará en cuenta el buen uso del espacio, lo conciso de su desarrollo y la presentación general de su informe, en particular de las figuras. La extensión recomendada es cinco páginas y no es necesario incluir portada (un encabezado basta) .

P1. Esta pregunta se enfoca en evaluar si, en promedio, el coeficiente intelectual (CI) de un chileno es superior 100.

Para este fin, comenzamos por denotar el CI de un chileno elegido aleatoriamente como X. Luego, asumiremos (de forma muy poco realista) que X es normalmente distribuido con media μ desconocida y desviación estándar conocida e igual a 16. Luego, consideremos una muestra aleatoria de $n = 16$ habitantes y denotaremos las hipótesis nula y alternativa respectivamente

$$H_0 : \mu = 100 \quad \& \quad H_A : \mu \geq 100. \tag{1}$$

A menos que se especifique lo contrario, consideramos un nivel de significancia $\alpha = 0.05$.

(a) Denote sus observaciones por x_i, enuncie el Z-test correspondiente y determine la región crítica

(b) Grafique la distribución del pivote Z e identifique la región crítica en su gráfico

(c) Asuma que el CI promedio (real) es $\mu = 110$ y grafique, sobre el mismo gráfico anterior, la distribución real. Calcule e indique en el gráfico los errores de Tipo I, Tipo II y la potencia del test.

(d) Denotando como μ la media real, calcule la potencia del test en función de μ y del número de muestras n. Grafique en función de ambas variables.

(e) Considere $n = 16$ y grafique la potencia vs μ para distintos valores de α. Analice los gráficos en términos del *trade off* entre errores de Tipo I y Tipo II.

P2. Consideremos las ventas de un producto, donde tenemos $i = 1, \ldots, I$ tiendas monitoreadas durante $j = 1, \ldots, n$ semanas. Consideramos un modelo de Poisson para las ventas de la i-ésima tienda en la j-ésima semana con parámetro θ_i, es decir,

$$x_{ij}|\theta_i \sim Po(\theta_i) \text{ i.i.d.} \tag{2}$$

Adicionalmente, las variaciones entre tiendas pueden modelarse asumiendo que los parámetros $\{\theta_i\}_{i=1}^I$ son independientes e idénticamente distribuidos de acuerdo a una distribución Gamma[1] dada por

$$\theta_i|\mu \sim Ga(a, a\mu), \tag{3}$$

donde a es un parámetro conocido y μ es desconocido.

Finalmente, asumiremos también que x_{ij} son condicionalmente independientes de μ dado θ_i, y adoptaremos la siguiente notación:

$$\mathbf{x} = \{x_{ij}\}_{i,j=1}^{I,n}, \quad \boldsymbol{\theta} = \{\theta_i\}_{i=1}^I. \tag{4}$$

(a) Calcule $p(x|\boldsymbol{\theta}, \mu)$

(b) Muestre que, condicional a x y a μ, los θ_i's son independientes y distribuidos de acuerdo a

$$\theta_i|X, \mu \sim Ga(a + ny_i, a\mu + n), \tag{5}$$

donde y_i es función de los datos. Identifique y_i.

(c) Encuentre la verosimilitud marginal $p(x|\mu)$ integrando con respecto a $\boldsymbol{\theta}$. Identifique los términos que son constantes y los que son *de interés* (con respecto a μ).

(d) Asuma la distribución *a priori* para $\mu \sim Ga(r, s)$ y muestre que

$$p(\mu|\mathbf{x}) \propto \mu^{r+aI-1} e^{-s\mu} (n + a\mu)^{-q}, \tag{6}$$

donde $q = I(a + n\bar{x})$ y $\bar{x}$ es la media muestral de todos los elementos de $\{x_{ij}\}_{i,j=1}^{I,n}$.

P3. El modelo lineal relaciona una variable dependiente $y \in \mathbb{R}$ con una independiente $x \in \mathbb{R}^d$ mediante la relación[2]

$$y = \theta^\top x, \tag{7}$$

donde $\theta \in \mathbb{R}^d$ es el conocido como el parámetro del modelo. En base a un conjunto de observaciones de la forma $\{(x_i, y_i)\}_{i=1}^n$, existen distintos estimadores puntuales de θ, donde los tres más usuales están dados por

$$\text{mínimos cuadrados ordinarios:} \quad \theta_{\text{MCO}} = \text{argmín} \sum_{i=1}^n (y_i - \theta^\top x_i)^2 \tag{8}$$

$$\text{regresión de } \textit{ridge}: \quad \theta_{\text{RR}} = \text{argmín} \sum_{i=1}^n (y_i - \theta^\top x_i)^2 + \rho \sum_{i=1}^n \theta_i^2 \tag{9}$$

$$\text{LASSO}: \quad \theta_{\text{LASSO}} = \text{argmín} \sum_{i=1}^n (y_i - \theta^\top x_i)^2 + \rho \sum_{i=1}^n |\theta_i|. \tag{10}$$

Como podemos ver, cada uno de los tres enfoques considera funciones de costo para la estimación del parámetro que difieren en un término de penalización de θ: MCO simplemente minimiza el error

[1] Use la siguiente parametrización para la distribución Gamma: $Ga(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$

[2] Asumimos que x ya contiene una coordenada constante para evitar expresar explícitamente el sesgo, es decir, escribimos $y = \theta^\top x$ en vez de $y = \theta^\top x + w$.

cuadrático medio de la regresión lineal, mientras que RR y LASSO[3] además penalizan valores altos de θ mediante el uso de las norma L_2 y L_1 respectivamente.

El objetivo de esta parte es interpretar los tres modelos anteriores y entender cómo pueden ser construidos desde un enfoque bayesiano, es decir, eligiendo los *priors* apropiados. Para esto, considere que la variable dependiente e independiente están relacionadas por

$$y|\theta, x \sim \mathcal{N}(\theta^\top x, \sigma^2) \tag{11}$$

$$\theta \sim p_0(\theta), \tag{12}$$

es decir, una verosimilitud gaussiana con media lineal con un prior sobre θ denotado por p_0.

(a) En base a las observaciones (datos) $D = \{(x_i, y_i)\}_{i=1}^n$, calcule la *distribución posterior predictiva* de y dada por

$$p(y|x, D), \tag{13}$$

y muestre que el *predictor puntual* de y dado por la esperanza condicional $\mathbb{E}(y|x, D)$ solo depende de x y de la esperanza posterior de θ, $\mathbb{E}(\theta|D)$.

(b) Muestre que el considerar un estimador de máxima verosimilitud para θ es equivalente a elegir θ de acuerdo al criterio de mínimos cuadrados ordinarios (MCO). Adicionalmente, especifique cuál debe ser el *prior* a elegir tal que el estimador *máximo a posteriori* (MAP) de θ sea también equivalente al estimador MCO. Discuta las propiedades de dicho *prior*.

(c) Ahora determine cuáles deben ser los *priors* sobre θ, tal que la estimación MAP de θ sea igual a la solución de RR y LASSO.

(d) Interprete los modelos construidos, en base a los distribuciones *a priori* construidas, ¿cómo puede interpretar MCO, RR y LASSO? ¿Qué es posible decir sobre las características de θ que promueven dichos *priors*?

(e) Implemente MCO, LASSO y RR en la base de datos *Breast Cancer* y complemente su análisis en base a los valores de θ encontrados por cada método, en particular, ¿qué puede decir de la magnitud de los elementos de θ en cada método? Utilice SKlearn, el cual le permitirá implementar todos los modelos en unas pocas líneas, su código debería empezar de la siguiente forma:

```
from sklearn.datasets import load_breast_cancer
data = load_breast_cancer()
print(data.keys()) #1era es input y 2da es output
print(data.feature_names) #variables de entrada
# importar modelos
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Ridge
from sklearn.linear_model import Lasso

```

[3] Acrónimo de *Least Absolute Shrinkage and Selection Operator*.