Tarea Práctica N°1 - MA5204

José Miguel Cordero Juan Pablo Cabeza

Abstract

En este informe de la Tarea Práctica I, estudiaremos modelos probabilísticos, en particular de Navi-Bayes, el cual es un algoritmo de clasificación que utiliza el teorema de Bayes. En esta clasificación, comparamos los valores posteriores de una observación para cada clase posible. Específicamente, debido a que la probabilidad marginal es constante en estas comparaciones, comparamos los numeradores del posterior para cada clase. Además, se exploran distintos métodos de regresión: Mínimos Cuadrados Ordinarios (MCO), Regularización de Ridge y LASSO (por sus siglas en inglés). Se analiza la relación entre los *prior* bayesianos y las distintas regularizaciones. Finalmente, se implementan los métodos y se compara el ajuste y error de cada uno de los métodos tanto para la base de datos completa como dividiendo en conjuntos de entrenamiento y prueba.

1. P1. Modelo de Naïve-Bayes

1.1. (a)

Para implementar el algoritmo, encontramos las probabilidades condicionales para cada atributo. En particular, las de tipo numérico se asume que siguen una distribución Gaussiana $\mathcal{N}(\mu, \sigma^2)$,

$$P(A_i = b|C_k) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} \exp\left(-\frac{(b-\mu_{ik})^2}{2\sigma_{ik}^2}\right),$$

con μ_{ik} y σ_{ik} la media muestral y desv. estad., respect. del atributo A_i, cuando la clase es C_k .

1.2. (b)

Aplicando el método de Naive-Bayes a los datos 'data_golf_test.csv', obtenemos que la diferencia de los atributos de los datos de testeo (correspondiente a la variable 'Play'), vienen dados por la siguiente tabla, (recordemos que esto es para predecir la variable 'Play', que tiene valor 'Yes' o 'No', así que tomamos la diferencia de su valor, dependiendo del valor que toma)

	$Yes - No$
0	0.000034
1	0.0000058
2	0.000092
3	0.000032

Por lo tanto, según el clasificador, todos los datos de 'data_golf_test.csv', clasifican como 'Play'= Yes.

2. P2. MCO, LASSO y RR

2.1. (a)

Se incluye en el código fuente.

2.2. (b)

regresión	R2	RMSE	Norma L1	Norma L2
Linear	0.6062	0.5243	2.058	0.992
Ridge	0.6062	0.5243	2.056	0.991
Lasso	0.2852	0.9517	0.151	0.145

Table 1. Resultados de métricas de regresión para *California Housing* (dataset completo)

En base a los resultados R^2 y RMSE obtenidos para cada modelo de regresión, se observa que el modelo de regresión lineal clásico (Mínimos Cuadrados Ordinarios, MCO) es el que obtiene mejores resultados de R^2 y error cuadrático medio (RMSE). La regresión Ridge obtiene prácticamente los mismos resultados que la regresión lineal clásica, tanto en R^2 y RMSE como en la magnitud de los parámetros obtenidos. Observando

Es interesante la relación entre la regularización clásica como penalización de los parámetros en la función de costo y la elección de los *prior* bayesianos. Si tomamos Tomando una *prior* uniforme $\pi(\theta) = 1$ (impropia) llegamos a

$$\theta_{MAP} = \arg\max_{\theta} p(Y|\widetilde{X}, \theta)\pi(\theta) = \arg\max_{\theta} p(Y|\widetilde{X}, \theta) = \theta_{EMV}$$

Sabemos que el Estimador de Máxima Verosimilitud conduce a MCO, por lo que ambos caminos son equivalentes.

Por otro lado, Al tomar una *prior* normal $\pi(\theta) \sim \mathcal{N}(0, \frac{\sigma^2}{\rho} I_{d+1})$ obtendremos que la expresión $\theta_{MAP} = \arg\max_{\theta} p(Y|\widetilde{X}, \theta)\pi(\theta)$ es la multiplicación de dos

exponenciales. Según el enunciado, $p(Y|\widetilde{X},\theta) \sim \mathcal{N}(y;\theta^T\widetilde{x},\sigma^2)$, por lo que al multiplicar ambos términos el θ queda cuadrático (por definición de la distribución exponencial) y sumado, tal como en la regularización Ridge. Otra vez, son equivalentes ambos caminos.

Para el caso de Lasso, tomando una *prior* de Laplace $\pi(\theta_i) \sim Laplace(0, \frac{2\sigma^2}{\rho})$ llegamos de nuevo a la multiplicación de dos exponenciales (Laplace es una distribución exponencial), pero donde θ ya no queda cuadrático sino que se suma en módulo, de forma análoga a la regularización de Lasso.

Es posible concluir que cada *prior* "fuerza" una determinada regularización. Es decir, fijar el *prior* del parámetro en una determinada región (en nuestro caso, cerca del 0) es equivalente a penalizar los parámetros que se escapen de esa región. En nuestros resultados se observa que el *prior* influye en el tamaño final de los parámetros.

Para el caso de la regularización Ridge, se asume una distribución normal en torno al cero, lo que podríamos considerar "suave" (razón por la cual los parámetros son similares a MCO), pero en la regularización LASSO se debe utilizar una distribución de Laplace, que es exponencial y deja el parámetro mucho más concentrado que una normal, lo que se refleja en que los parámetros de LASSO sean mucho menores en norma L1 y L2 que los parámetros de Ridge.

La forma del *prior* de LASSO también permite explicar por qué LASSO "selecciona" variables: dado que el parámetro se asume más concentrado en torno al cero, va a ser más probable que sea 0 para variables poco relevantes, manteniendo aquellas variables que en una cantidad significativa aporten a la variable independiente.

2.3. (c)

regresión	R2 Train	R2 Test	RMSE Train	RMSE Test
Linear	0.6020	0.6223	0.528	0.508
Ridge	0.6020	0.6223	0.528	0.508
Lasso	0.2784	0.287	0.957	0.960

Table 2. Resultados de métricas de regresión para *California Housing* (80% train y 20% test)

Al separar en conjuntos de entrenamiento (80%) y prueba (20%), y entrenando con el conjunto de entrenamiento, se observa que tanto MCO como Ridge obtienen mejores resultados en prueba que en entrenamiento, dado que mejoran su R^2 y disminuye su RMSE. se puede concluir que ambos modelos generalizaron. Si bien LASSO consigue aumentar ligieramente su R^2, también aumenta su error, y su performance es mejor que los otros regresores.

Posibles explicaciones de este resultado se observan en los gráficos de magnitudes de los parámetros de cada atributo. LASSO apenas captura dos atributos, en cuanto MCO y Ridge capturan la mayoría de los atributos, por lo que la selección realizada por LASSO (forzada a través de la regularización l_ρ $\rho = 1$) le está quitando capadidad al modelo de explicar la variable independiente.