Función de Coste

Tópicos de Investigación : Machine Learning

Josué Lucero Rivera

Universidad Nacional de Ingeniería

29 de Setiembre

Deseamos encontrar una función: $f : X \to \mathrm{R}$ tal que para $(x, y) \in X \times Y$ el valor $f(x)$ es una buena predicción de y a x. La siguiente definición nos ayudará a definir lo que entendemos por *buena predicción*.

Definición:

Sea $(X, \mathcal{A})$ un espacio medible, y $Y \subset \mathrm{R}$ un subconjunto cerrado. Entonces la función $L : X \times Y \times \mathrm{R} \to [0, \infty\rangle$ es llamada una **función de coste**, o simplemente de **coste**, si es medible.

Interpretaremos $L(x, y, f(x))$ como el *costo*, de predicción y por $f(x)$ si x es observado, es decir, cuanto menor es el valor $L(x, y, f(x))$, $f(x)$ que predice y en el sentido de L.

Recordemos ahora desde la introducción que nuestro principal objetivo es tener un Coste medio de las observaciones futuras (x, y) que no se ven. Esto conduce a la siguiente definición.

Definición:

Sea $L : X \times Y \times \mathrm{R} \to [0, \infty\rangle$ una función de Coste, y P una medida de probabilidad en $X \times Y$. Entonces, para una función medible $f : X \to \mathrm{R}$, el L-**riesgo** es definido por:

$$R_{L,P}(f) = \int_{X \times Y} L(x, y, f(x)) dP(x, y)$$

$$= \int_X \int_Y L(x, y, f(x)) dP(y|x) dP_X(x)$$

Para una sucesión dada $D = ((x_1, y_1), \ldots, (x_n, y_n) \in (X \times Y)^n$, escribimos $D = \frac{1}{n} \sum_{i=1}^{n} \delta_{(x_i, y_i)}$, donde $\delta_{(x_i, y_i)}$ es la medida de Dirac en (x_i, y_i).

En otras palabras, D es la medida empírica asociada a D.

El riesgo de la función $f : X \to \mathrm{R}$ con respecto a esta medida es llamado el L-**riesgo empírico**

$$R_{L,D}(f) = \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, f(x_i))$$

Definición:

Sea $L : X \times Y \times \mathrm{R} \to [0, \infty\rangle$ una función de Coste y P una medida de probabilidad en $X \times Y$. El L-riesgo minimal

$$R^*_{L,P} = \inf\{R_{L,P}(f) \mid f : X \to \mathrm{R} \textit{ es medible}\}$$

es llamado el `riesgo de Bayes` con respecto a P y L.

Además, una función medible $f^*_{L,P} : x \to \mathrm{R}$ con $R_{L,P}(f^*_{L,P})$ es llamada una **función de decisión de Bayes**.

Por lo general, el primer paso para resolver un problema de aprendizaje práctico es encontrar una función de de Coste que mejor describe el aprendizaje.

En general, la elección de una función de Coste adecuado depende de una aplicación específica. Sin embargo, hay algunos escenarios básicos de aprendizaje que a menudo se ajustan al problema de aprendizaje,

Veamos unos ejemplos:

Clasificación Binaria

Sean $Y = \{-1, 1\}$ y P una distribución que genera datos desconocidos en $X \times Y$. Entonces, la meta en clasificación binaria es predecir y del par (x, y) trazada desde P si solo x es observado.

La función de Coste más común que describe esta meta de aprendizaje es la **clasificación de de Coste**

$L_{class} : Y \times \mathrm{R} \to [0, \infty\rangle$, que es definida por

$$L_{class}(y, t) = \mathbf{1}_{\langle -\infty, \mathbf{0}]}(y \ sign \ t), \qquad y \in Y, t \in \mathrm{R}$$

donde utilizamos la convención sign $0{:=}1$.

Notamos que L_{class} solo penaliza predicciones t cuyo signos son contrarios a los de y, por lo que de hecho refleja nuestro objetivo de aprendizaje. Ahora, sea una función medible $f : X \to \mathrm{R}$, unos cálculos simples nos muestran:

$$\begin{aligned} R_{L_{class},P}(f) &= \int_X \eta(x)\mathbf{1}_{\langle -\infty,\mathbf{0}]}(f(x)) + \mathbf{1}_{\langle \mathbf{0},-\infty]}(f(x))dP_X(x) \\ &= P(\{(x,y) \in X \times Y \ \textit{sign}\ f(x) \neq y\}) \end{aligned}$$

donde $\eta(x) := P(y = 1|x)$, $x \in X$.

De esto podemos concluir que f es una función de decisión de Bayes si y sólo si $(2\eta(x) - 1)$ sign $f(x) \geq 0$ para P_X c.t.p. $x \in X$. Además esta consideración nos da

$$R^*_{L_{class},P} = \int_X \min\{\eta, 1 - \eta\} dP_X$$

Observaciones :

La función de Coste L_{class} iguala los pesos de ambos tipos de errores, a saber $y = 1$ cuando $f(x) < 0$, y $y = -1$ cuando $f(x) \geq 0$. Esto tiene sentido especialmente en situaciones en las que se desea clasificar objetos tales como caracteres escritos a mano o imágenes.

En muchas situaciones prácticas, sin embargo, ambos tipos de error deben ser objeto de ponderaciones diferentes. Por ejemplo, si se quiere detectar intrusiones en la red informática, en función de los recursos disponibles para la investigación de las alarmas y la sensibilidad de la red, los dos tipos de errores es probable que tengan diferentes costes reales.

Clasificación Binaria Ponderada

Sea $Y = -1, 1$ y $\alpha \in \langle 0, 1 \rangle$. Entonces la α-**de Coste de clasificación ponderada** $L_{\alpha-class} : Y \times \mathrm{R} \to [0, \infty\rangle$ es definida por:

$$L_{\alpha-class} = \begin{cases} 1 - \alpha, & \text{si } y = -1, t < 0 \\ \alpha, & \text{if } y = -1, t \geq 0 \\ 0, & \text{en otro caso.} \end{cases}$$

para todo $y \in Y$, $t \in \mathrm{R}$.

Es evidente que tenemos $2L_{\frac{1}{2}-class} = L_{class}$, es decir, la clasificación binaria estándar es un caso especial de la clasificación general ponderada.

Ahora, tenemos la medida de probabilidad P en $X \times Y$ y un medible $f : X \to \mathrm{R}$, el $L_{\alpha-class}$-riesgo puede ser calculado por:

$$R_{L_{\alpha-class},P}(f) = (1-\alpha)\int_{f<0} \eta dP_X + \alpha \int_{f\geq 0} (1-\eta) dP_X$$

Dónde $\eta(x) = P(y = 1|x)$, $x \in X$.

De esto concluimos que f es una función de decisión de Bayes si y sólo si $(\eta(x) - \alpha)sign\ f(x) \geq 0$ para P_X- casi todo punto $x \in X$.

Finalmente, el $L_{\alpha-class}$-riesgo de Bayes es

$$R^*_{L_{\alpha-class},P} = \int_X \min\{(1-\alpha)\eta, \alpha(1-\eta)\} dP_X$$

En este ejemplo la meta es predecir y a partir del conjunto $\{-1, 1\}$.

Regresión de mínimos cuadrados

La meta en regresión es predecir $y \in Y = \mathrm{R}$ del par (x, y) a partir de una medida de probabilidad desconocida P en $X \times Y$ si observamos solo x.

La forma más común de formalizar esta meta está basada en la fórmula **de Coste mínima cuadrada** $L_{LS} : Y \times \mathrm{R} \to [0, \infty\rangle$ definida por

$$L_{LS}(y, t) = (y - t)^2, \qquad y \in Y, t \in \mathrm{R}$$

En otras palabras, la fórmula de Coste mínima cuadrada minimiza el coste cuadrático entre y y t. Evidentemente para cada función medible $f : X \to \mathrm{R}$, el L_{LS}-riesgo es

$$R_{L_{LS},P} = \int_X \int_Y (y - f(x))^2 dP(y|x) dP_X(x)$$

y sólo si $f(x)$ es casi igual a Y evaluado en x, es decir, si y sólo si:

$$f(x) = \mathrm{E}_P(Y|x) = \int_Y dP(y|x)$$

Para P_X en casi todo punto $x \in X$.

Para minimizar la integral interior respecto de $f(x)$, tenemos que tener en cuenta que f es una función de decisión de Bayes si y sólo si $f(x)$ es casi igual a Y evaluado en x, es decir, si y sólo si:

$$f(x) = \mathrm{E}_P(Y|x) = \int_Y dP(y|x)$$

Para P_X en casi todo punto $x \in X$.

Además, si hacemos $x \mapsto \mathrm{E}_P(Y|x)$ en $R_{L_{LS},P}(\cdot)$ muestra que el L_{LS}-riesgo de Bayes es el promedio condicional Y- varianza, es decir,

$$R^*_{L_{LS},P} = \int_X \mathrm{E}_P(Y^2|x) - (\mathrm{E}_P(Y|x))^2 dP_x(x)$$

Finalmente, un cáculo básico muestra que el L_{LS}-riesgo de *exceso* de $f : X \to \mathrm{R}$ es

$$R_{L_{LS}}(f) - R^*_{L_{LS},P} = \int_X (\mathrm{E}_P(Y|x) - f(x))^2 dP_X(x)$$

Es decir, si $R_{L_{LS},P}(f)$ está cerca de $R^*_{L_{LS},P}$, entonces f está cerca de una función de decisión de Bayes en el sentido de la $\|\cdot\|_{L_2(P_X)}$

Bibliografía

Duchi, Jhon

Supplemental Lecture notes.

Stanford CS 229

Berger, James

Statistical Decision Theory and Bayesian Analysis. 2th ed.

Springer Verlag