Trabajo práctico 1: Selección de Características Estadística

Ph. D. Saúl Calderón Ramírez
Instituto Tecnológico de Costa Rica,
Escuela de Computación
PAttern Recongition and MAchine Learning Group (PARMA-Group)

2 de octubre de 2023

Fecha de entrega: Lunes 16 de Octubre.
Entrega: Un archivo .zip con el código fuente LaTeX o Lyx, el pdf, y un jupyter en Pytorch, debidamente documentado, con una función definida por ejercicio. A través del TEC-digital.
Modo de trabajo: Grupos de 3 personas.

Resumen

En el presente trabajo práctico se introduce un mecanismo simple de selección de características usando la distancia entre las densidades de las características.

1. Selección de características usando la distancia entre las densidades

La selección de características para un arreglo de entrada $\overrightarrow{x} \in \mathbb{R}^D$ consiste en implementar una transformación $\overrightarrow{x}' = T(\overrightarrow{x})$ la cual genere un nuevo arreglo $\overrightarrow{x}' \in \mathbb{R}^{D'}$. La seleccion de caracteristicas selecciona las dimensiones del arreglo $\overrightarrow{x}$ las cuales contengan la mayor información para el problema de interés. En el contexto de la clasificación supervisada de una observación $\overrightarrow{x}$ en por ejemplo $K = 2$ categorias (por lo que entonces $t_i \in \{0, 1\}$), para la cual se cuenta entonces con un conjunto de observaciones

$$X = \begin{bmatrix} - & \overrightarrow{x}^{(1)} & - \\ - & \overrightarrow{x}^{(2)} & - \\ & \vdots & \\ - & \overrightarrow{x}^{(N)} & - \end{bmatrix} \qquad \overrightarrow{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$$

es posible implementar un criterio de que tan «discriminatorio» es una dimensión o característica x_d, usando una idea simple: medir que tan diferentes son

las densidades de esa caracteristica de los datos que pertenecen a la clase 1 y los datos que pertenecen a la clase 0. Si la dimension es muy «discriminatoria», entonces la distancia entre ambas densidades debe ser alta, de lo contrario Para seguir esta idea, podemos desarrollar los siguientes pasos:

1. Estimar las densidades de los datos que pertenecen a la clase 0 y a la clase 1, para cada caracteristica o dimension de entrada d, generando una lista de pares ordenados de densidades:

$$\begin{matrix} \left\langle \overrightarrow{p}_1^{(0)}, \overrightarrow{p}_1^{(1)} \right\rangle \\ \left\langle \overrightarrow{p}_2^{(0)}, \overrightarrow{p}_2^{(1)} \right\rangle \\ \vdots \\ \left\langle \overrightarrow{p}_D^{(0)}, \overrightarrow{p}_D^{(1)} \right\rangle \end{matrix}$$

   Para estimar las densidades, puede calcularse el histograma de cada caracteristica, tomando en cuenta claro la pertenencia a cada clase.

2. Para comparar las estimaciones de las funciones de densidad de probabilidad puede usarse la divergencia de Kullback-Leibler por ejemplo. Esta divergencia, compara dos densidades $\overrightarrow{p}$ y $\overrightarrow{q}$ como sigue:

$$d_{\mathrm{KL}}\left(\overrightarrow{p}, \overrightarrow{q}\right) = \sum_i^L p_i \cdot \log\left(\frac{p_i}{q_i}\right)$$

   donde L es la cantidad de cubetas o eventos posibles para tal dimension. La Figura 1 muestra el concepto de comparacion de dos funciones de densidad de probabilidad. Esta divergencia (o cualquier otra que compare dos funciones de densidad de probabilidad), sirve entonces como un *puntaje* para estimar que dimensiones son mas valiosas de retener o no. Por cada par de densidades del punto anterior, calcule entonces la divergencia de Kulback-Leibler, generando entonces una lista de puntajes:

$$\begin{matrix} d_{\mathrm{KL}}\left(\overrightarrow{p}_1^{(0)}, \overrightarrow{p}_1^{(1)}\right) \\ d_{\mathrm{KL}}\left(\overrightarrow{p}_2^{(0)}, \overrightarrow{p}_2^{(1)}\right) \\ \vdots \\ d_{\mathrm{KL}}\left(\overrightarrow{p}_D^{(0)}, \overrightarrow{p}_D^{(1)}\right) \end{matrix}$$

3. Al ordenar tal lista de puntajes, de mayor a menor, es posible elegir una cantidad τ de dimensiones o caracteristicas. De esta forma, el nuevo espacio de caracteristicas quedara conformado por τ dimensiones.

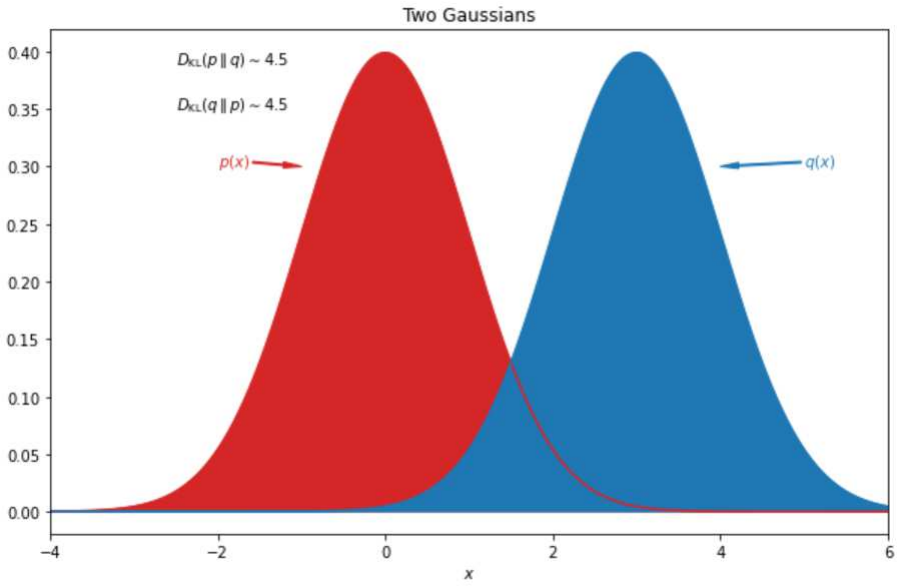

Figura 1: Comparación de dos funciones de densidad de probabilidad.

2. Implementación de la selección de características usando la divergencia Kullback-Leibler

Para el conjunto de datos definido en el archivo *pimaindiansdiabetes*.csv, tomado de `https://www.kaggle.com/kumargh/pimaindiansdiabetescsv`, el cual contiene la siguiente información de alrededor de 768 pacientes, etiquetados de forma binaria por si sufrieron o no diabetes en los proximos 5 años luego de la toma de los datos:

1. x_0: Number of times pregnant.
2. x_1 : Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
3. x_2 : Diastolic blood pressure (mm Hg).
4. x_3 : Triceps skinfold thickness (mm).
5. x_4 : 2-Hour serum insulin (mu U/ml).
6. x_5 : Body mass index (weight in kg/(height in m)^2).
7. x_6 : Diabetes pedigree function.
8. x_7 : Age (years).

El archivo base lee un *dataframe* con *pandas* tal conjunto de datos. Use pytorch para resolver todos los siguientes ejercicios:

1. **(40 puntos)** Implemente de la forma mas vectorial posible (prescindiendo al maximo de ciclos *for*) el primer paso del proceso descrito en la sección anterior. Guarde en un solo tensor la matriz con las densidades resultantes.

    a) Grafique las densidades para cada *feature,* y comente los resultados.

2. (**20 puntos**) Implemente de la forma mas vectorial posible (prescindiendo al maximo de ciclos *for*) la funcion *calculate_kl_divergence(p_densities_1, p_densities_2)* la cual tome dos matrices con D densidades, y las compare usando la divergencia KL definida en la seccion anterior. Tal funcion debe retonar la lista de divergencias KL para todos los par de densidades recibidos, en un tensor.

    a) Indique el valor de divergencia KL por para cada gráfica del apartado anterior.

3. (**20 puntos**) Implemente la funcion *select_best_tau_features_from_kl_list* la cual ordene la lista de puntajes obtenida en el punto anterior y seleccione los τ mejores dimensiones a partir del conjunto de datos original X. La funcion debe retornar la matriz X' con las observaciones transformadas.

    a) Implemente la funcion *select_best_features_kl* la cual tome el conjunto de datos original X y la cantidad de dimensiones a preservar τ y realice todos los pasos anteriores.

4. (**20 puntos**) Experimente preservando $\tau = 2, 3, 4, 5$ dimensiones, y compare el comportamiento del sistema con no eliminar ninguna dimension. Como metrica, utilice la tasa de aciertos del algoritmo del algoritmo SVM implementado en el codigo base usando una particion de test especifica. Reporte los resultados en una tabla y comentelos con detalle.