Trabajo de fin de grado

El aprendizaje de métricas de distancia: análisis y revisión de técnicas desarrolladas en alta dimensionalidad

Juan Luis Suárez Díaz
Universidad de Granada
jlsuarezdiaz@correo.ugr.es

Índice

Parte I

Matemáticas

Parte II

Informática teórica

El aprendizaje automático

El aprendizaje de métricas de distancia

Descripción teórica de las técnicas de aprendizaje de métricas de distancia

1. Técnicas de reducción de dimensionalidad

1.1. PCA

TODO

1.2. LDA

TODO

1.3. ANMM

TODO

2. Técnicas orientadas a la mejora del clasificador de vecinos cercanos

2.1. LMNN

LMNN (*Large Margin Nearest Neighbors*) [1] es un algoritmo de aprendizaje de métricas de distancia orientado específicamente a mejorar la precisión del clasificador kNN. Se basa en la premisa de que el kNN clasificará con más fiabilidad un ejemplo si sus k vecinos comparten la misma etiqueta, y para ello intenta aprender una distancia que maximice el número de ejemplos que comparten etiqueta con el mayor número de vecinos posible.

De esta forma, el algoritmo LMNN trata de minimizar una función de error que penaliza, por un lado, las distancias grandes entre cada ejemplo y los considerados como sus vecinos ideales, y por otro lado, las distancias pequeñas entre ejemplos de distintas clases.

Supongamos que tenemos un conjunto de datos $\mathcal{X} = \{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ con etiquetas $\mathcal{Y} = \{y_1, \ldots, y_d\}$. Para su funcionamiento, el algoritmo hace uso del concepto de *vecinos objetivo* o *target neighbors*. Dado un ejemplo $x_i \in \mathcal{X}$, sus k vecinos objetivos son aquellos ejemplos de la misma clase que x_i, y distintos de este, para los que se desea que sean considerados como vecinos en la clasificación del kNN. Si x_j es un vecino objetivo de x_i, entonces lo notaremos $j \to i$. Estos vecinos objetivo están fijos durante el proceso de aprendizaje. Si se dispone de alguna información a priori se puede utilizar para determinarlos. En caso contrario, una buena opción es utilizar los vecinos cercanos de la misma clase para la distancia euclídea.

Una vez establecidos los vecinos objetivo, para cada distancia y para cada ejemplo que manejemos podemos establecer un perímetro determinado por el vecino más lejano a dicho ejemplo. Buscamos distancias para las cuales no haya ejemplos de otras clases en dicho perímetro. Hay que destacar que con este perímetro no hay suficientes garantías de separación, pues la distancia encontrada podría haber colapsado todos los vecinos objetivo en un punto y entonces el perímetro tendría radio cero. Por ello se considera un margen determinado por el radio del perímetro, al que se añade una constante positiva. Veremos que no hay pérdida de generalidad, por la función objetivo que vamos a definir, en suponer que dicha constante es 1. A cualquier ejemplo de distinta clase que invada este margen lo llamaremos *impostor*. Nuestro objetivo, por tanto, será, además de acercar cada ejemplo a sus vecinos objetivo lo máximo posible, intentar alejar lo máximo posible a los impostores.

TODO

2.2. NCA

TODO

3. Técnicas orientadas a la mejora del clasificador de centroides cercanos

3.1. NCMML

TODO

3.2. NCMC

TODO

4. Técnicas basadas en teoría de la información

4.1. ITML

TODO

4.2. DMLMJ

TODO

4.3. MCML

TODO

5. Otras técnicas de aprendizaje de métricas de distancia

5.1. LSI

TODO

5.2. DML-eig

TODO

5.3. LDML

TODO

6. El kernel trick. Algoritmos de aprendizaje de métricas de distancia basados en kernels

6.1. El kernel trick

TODO

6.2. KLMNN

TODO

6.3. KANMM

TODO

6.4. KDMLMJ

TODO

6.5. KDA

TODO

Referencias

[1] Kilian Q Weinberger y Lawrence K Saul. "Distance metric learning for large margin nearest neighbor classification". En: *Journal of Machine Learning Research* 10.Feb (2009), págs. 207-244.