Manifold Learning

Mateo Dulce Rubio

Mayo 2017

Referencia principal:

Semi-Supervised Learning on Riemmanian Manifolds,
Mikhail Belkin & Partha Niyogi, 2004.

Problema de clasificación

Supongamos que tenemos un espacio $\mathcal{X}$, cuyos elementos debemos clasificar entre las dos clases $C_1,\ C_2$. El modelo probabilistico para este problema es la función de densidad de probabilidad $p(x)$, y las densidades de cada clase $\{p(C_1|x \in \mathcal{X})\},\ \{p(C_2|x \in \mathcal{X})\}$.

Es decir, necesitamos conocer cómo se distribuyen los datos, y como se asignan a cada una de las clases.

Motivación

Los datos no etiquetados (*unlabeled data*) por sí mismos no dan mucha inforamación sobre las reglas de asignación a las clases, justamente por no tener etiqueta.

Sin embargo, es difícil (costoso) conseguir datos etiquetados (*label data*).

Motivación

¿Cómo con relativamente pócos datos etiquetados podemos clasificar muchos datos no etiquetados?

$\rightarrow$ *Semi-Supervised Machine Learning*

Aproximación

Explotar la estructura intrínseca de los datos para mejorar la clasificación de datos no etiquetados, bajo el supuesto de que los datos residen en una variedad de baja dimensión dentro de un espacio de representación de alta dimensión.

Aproximación

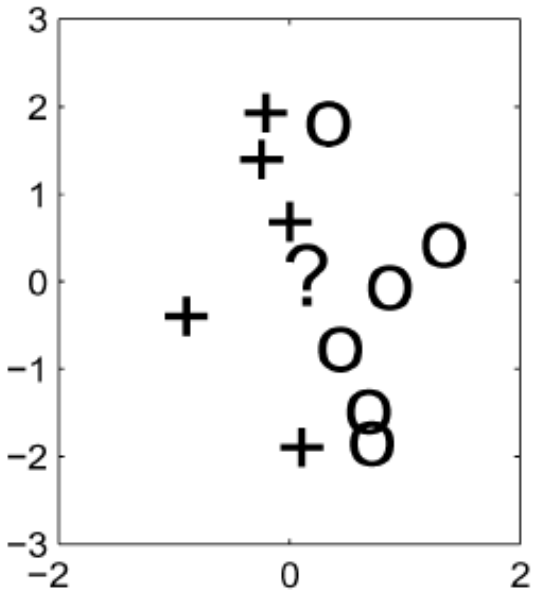
Por ejemplo, en clasificación de texto, los documentos se representan usualmente por vectores cuyos elementos son contadores de palabras en el documento. Aunque no está claro por qué el espacio de documentos debería ser una variedad, sí se puede ver que tiene una estructura compleja y que solo ocupa una pequeña parte dentro del gigantesco espacio de representación.

Aproximación

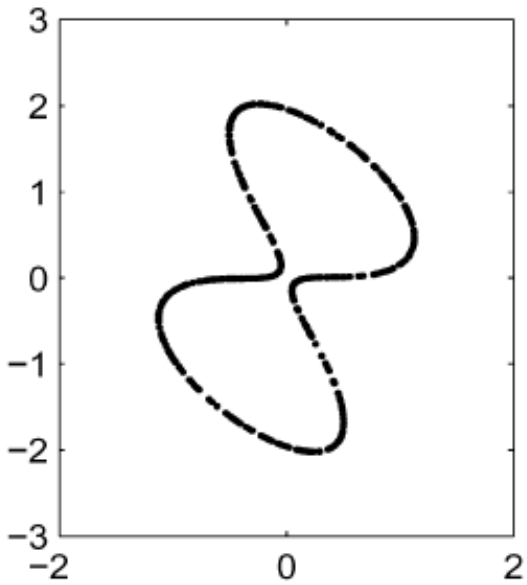
Los datos no etiquetados no dan información sobre la regla de asignación, pero sí sobre la distribución $(p(x))$ de los datos en el espacio.

Por su parte, los datos etiquetados sí nos permiten estimar cómo se clasifican los datos en el espacio, $\{p(C_1|x \in \mathcal{X})\}$, $\{p(C_2|x \in \mathcal{X})\}$.

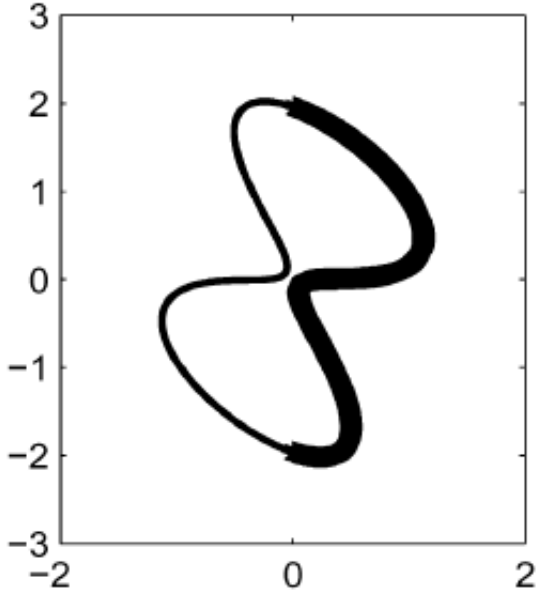
Ejemplo



Labeled data



Unlabeled data



Underlying manifold

Ejemplo

- No siempre se puede clasificar ? solo con los datos etiquetados.

Ejemplo

- No siempre se puede clasificar ? solo con los datos etiquetados.
- Los datos no etiquetados no dan información sobre las reglas de clasificación, pero sí del espacio de representación, que asumimos es una variedad.

Ejemplo

- Es más importante la distancia (geodésica) sobre la variedad subyacente, que la distancia Euclideana del plano.

Ejemplo

- Es más importante la distancia (geodésica) sobre la variedad subyacente, que la distancia Euclideana del plano.
- Puntos que inicialmente parecen muy cercanos en el plano, estan a lados opuestos de nuestra variedad.

Ejemplo

- Es más importante la distancia (geodésica) sobre la variedad subyacente, que la distancia Euclideana del plano.
- Puntos que inicialmente parecen muy cercanos en el plano, estan a lados opuestos de nuestra variedad.
- Es por esto que se deben crear los clasificadores sobre la variedad subyacente, y no sobre el espacio de representación completo.

Ejemplo

- Aunque los datos sugieren una variedad subyacente, el problema no es completamente trivial dado que dos partes de la curva se acercan demasiado, y puede causar confusión

Ejemplo

- Aunque los datos sugieren una variedad subyacente, el problema no es completamente trivial dado que dos partes de la curva se acercan demasiado, y puede causar confusión
- Muchas formas de representar la misma variedad (homeomorfismos, bajo cambio de coordenadas).

Problema de Clasificación

Recapitulando:

- Recuperar la variedad en la que viven los datos, y desarrollar clasificadores sobre esta nos da una ventaja en el Problema de Clasificación inicial.

Problema de Clasificación

Recapitulando:

- Recuperar la variedad en la que viven los datos, y desarrollar clasificadores sobre esta nos da una ventaja en el Problema de Clasificación inicial.
- Los datos no etiquetados nos sirven para recuperar la variedad, mientras con los datos etiquetados desarrollamos clasificadores sobre dicha variedad.

Problema de Clasificación

Mas formalmente:

Dado un conjunto de datos etiquetados $((x_i, y_i) : x_i \in \mathbb{R}^k, y_i \in Y)$, y un conjunto de datos no etiquetados $(x_j \in \mathbb{R}^k)$, queremos encontrar un clasificador

$$f : \mathbb{R}^k \to Y.$$

Como k puede ser demasiado grande, nos enfrentamos con "the curse of dimensionality".

Problema de Clasificación

Pero, si aprovechamos el hecho que los datos viven en una subvariedad $\mathcal{M}$ del espacio de representación $\mathbb{R}^k$, de menor dimensión, reducimos el problema a encontrar un clasificador

$$f : \mathcal{M} \to Y, \qquad x_k \in \mathcal{M} \subset \mathbb{R}^k, \quad dim(\mathcal{M}) < k.$$

Aproximación

1. Representar los datos como una variedad:
   - Tomando los datos como vértices, contruímos una grafo ponderado.

Aproximación

1. Representar los datos como una variedad:
   - Tomando los datos como vértices, contruímos una grafo ponderado.
   - Dos puntos (datos) están conectados si, y solo si, son adjacentes en algún sentido (e.g. estan a distancia menor que algún ϵ, o pertenecen al conjunto de n vecinos más cercanos).

Aproximación

1. Representar los datos como una variedad:
   - Tomando los datos como vértices, contruímos una grafo ponderado.
   - Dos puntos (datos) están conectados si, y solo si, son adjacentes en algún sentido (e.g. estan a distancia menor que algún ϵ, o pertenecen al conjunto de n vecinos más cercanos).
   - A cada arco se le asocia una distancia entre los dos vertices que relaciona.

Aproximación

1. Representar los datos como una variedad:
   - A cada par de datos se les puede asociar "distancia geodésica", definida como la longitud del camino más corto que los une.

Aproximación

1. Representar los datos como una variedad:
   - A cada par de datos se les puede asociar "distancia geodésica", definida como la longitud del camino más corto que los une.
   - La distancia geodésica puede variar respecto a la distancia natural del espacio ambiente.

Aproximación

1. Representar los datos como una variedad:
   - A cada par de datos se les puede asociar "distancia geodésica", definida como la longitud del camino más corto que los une.
   - La distancia geodésica puede variar respecto a la distancia natural del espacio ambiente.
   - Pero se puede mostrar que si los datos se ditribuyen con una distribución de probabilidad con soporte sobre toda la variedad de representación, la distancia geodésica estimada tiende a la distancia geodésica actual de la variedad ambiente, cuando el número de datos tiende a infinito.

Aproximación

2. Estimación función de clasificacion:
   - Un proceso natural sería usar la distancia geodésica definida sobre la variedad, para construir "vecinos geodésicos más cercanos".

Aproximación

2. Estimación función de clasificacion:
   - Un proceso natural sería usar la distancia geodésica definida sobre la variedad, para construir "vecinos geodésicos más cercanos".
   - Si para un punto no etiquetado u, se tiene que el punto etiquetado l es su vecino geodésico más cercano (la distancia sobre los arcos es la menor posible), entonces la etiqueta de l se le asigna a u.

Aproximación

2. Estimación función de clasificacion:
   - Un proceso natural sería usar la distancia geodésica definida sobre la variedad, para construir "vecinos geodésicos más cercanos".
   - Si para un punto no etiquetado u, se tiene que el punto etiquetado l es su vecino geodésico más cercano (la distancia sobre los arcos es la menor posible), entonces la etiqueta de l se le asigna a u.
   - Solución inestable, y sensible a ruido o valores atípicos.

Aproximación

2. Estimación función de clasificación:
   - Basada en el operador de Laplace-Baltrami sobre la variedad, Δ.

Aproximación

2. Estimación función de clasificación:
   - Basada en el operador de Laplace-Baltrami sobre la variedad, Δ.
   - Si $\mathcal{M}$ es una variedad compacta, Δ tiene un espectro discreto, y las funciones propias de Δ son una base ortogonal para el espacio de Hilbert $\mathcal{L}^2(\mathcal{M})$.

Aproximación

2. Estimación función de clasificación:
   - Basada en el operador de Laplace-Baltrami sobre la variedad, Δ.
   - Si $\mathcal{M}$ es una variedad compacta, Δ tiene un espectro discreto, y las funciones propias de Δ son una base ortogonal para el espacio de Hilbert $\mathcal{L}^2(\mathcal{M})$.
   - Por lo tanto, toda función $f \in \mathcal{L}^2(\mathcal{M})$ puede escribirse como

$$f(x) = \sum_{i=0}^{\infty} a_i e_i(x),$$

   con e_i son funciones propias: $\Delta e_i = \lambda_i e_i$.

Aproximación

2. Estimación función de clasificación:
   - De esta manera, asumiendo que los datos viven en una variedad compacta $\mathcal{M}$, la función de clasificación puede ser representada por una función cuadrado-integrable

$$m : \mathcal{M} \to \{-1, 1\}.$$

   - Solo neceistamos que $m(x)$ sea una función medible.

Aproximación

2. Estimación función de clasificación:
   - Podemos interpretar el Problema de Clasificación como un problema de interpolación de una función sobre una variedad.

Aproximación

2. Estimación función de clasificación:
   - Podemos interpretar el Problema de Clasificación como un problema de interpolación de una función sobre una variedad.
   - Como toda función se puede escribir en términos de las funciones propias del Laplaciano, solo debemos ajustar los coeficientes para que concuerden con los datos etiquetados.

$$m(x) \approx \sum_{i=0}^{N} a_i e_i(x).$$

Aproximación

2. Estimación función de clasificación:
   - Podemos interpretar el Problema de Clasificación como un problema de interpolación de una función sobre una variedad.
   - Como toda función se puede escribir en términos de las funciones propias del Laplaciano, solo debemos ajustar los coeficientes para que concuerden con los datos etiquetados.

$$m(x) \approx \sum_{i=0}^{N} a_i e_i(x).$$

   - Las funciones propias del Laplaciano no son solo una base natural a considerar, sino que además satisfacen una condición de optimalidad, en el sentido que proveen la aproximación más suave.