Práctica 1
Mixturas de Gaussianas

Vicente Gras Mas

Introducción

En esta práctica probaremos diferentes técnicas para clasificar datos mediante aprendizaje automático.
La base de datos que vamos a usar es la MNIST, que consiste en una colección de imágenes de dígitos manuscritos (del 0 al 9, por tanto 10 clases). Concretamente dispondremos para las diferentes pruebas, 60000 imágenes de entrenamiento y 10000 de test.
A continuación se harán varias pruebas con un clasificador multinomial el cual definirá una frontera de decisión lineal (Ejercicio 4), y luego implementaremos un clasificador que defina una tasa de error cuadrática como es un clasificador gaussiano (Ejercicio 5) para ver la diferencia de error de clasificación entre un clasificador lineal y uno cuadrático.

Ejercicio 4.1

Realiza un experimento para evaluar el error del clasificador gaussiano(sin suavizado, es decir,α= 1,0) en función del número de componentes PCA a las cuales se proyectan los datos originales. Representa gráficamente los resultados obtenidos en forma de una curva.

Para resolver este ejercicio lanzaremos el script *ejercicio4_1.m* que nos devolverá el archivo *ejercicio4_1.dat* , una vez habiendo ejecutado este paso, entrando en el gnuplot en una terminal, lanzando la orden load ‘plot_4_1.plt’ (donde ‘plot_4_1.plt’ es un script gnuplot) obtendremos la gráfica que representa lo pedido en el ejercicio 4.1(figura 1):

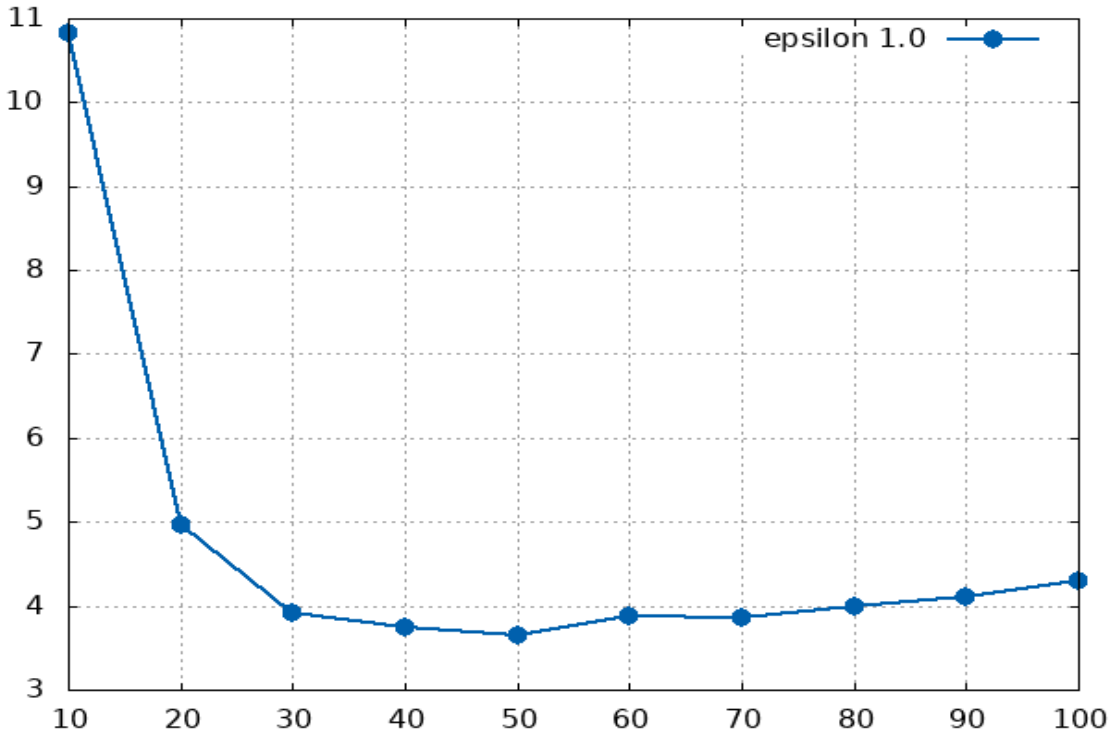


figura 1

Ejercicio 4.2

A la vista de los resultados obtenidos en el ejercicio anterior, repite el experimento anterior ajustando el factor de suavizado. A ñade a la gráfica del ejercicio anterior una curva de error de clasificación en función del número de componentes PCA por cada valor α utilizado. Compara los resultados obtenidos con los reportados en la tarea MNIST.

Usando el script ejercicio4_2.m, el cual contiene el mismo código que el ejercicio anterior añadiendo un bucle interno para variar α, y usando el archivo devuelto (ejercicio_4_ 2.dat) en nuestro script plot_4_ 2.plt obtenemos la siguiente gráfica (figura 2):

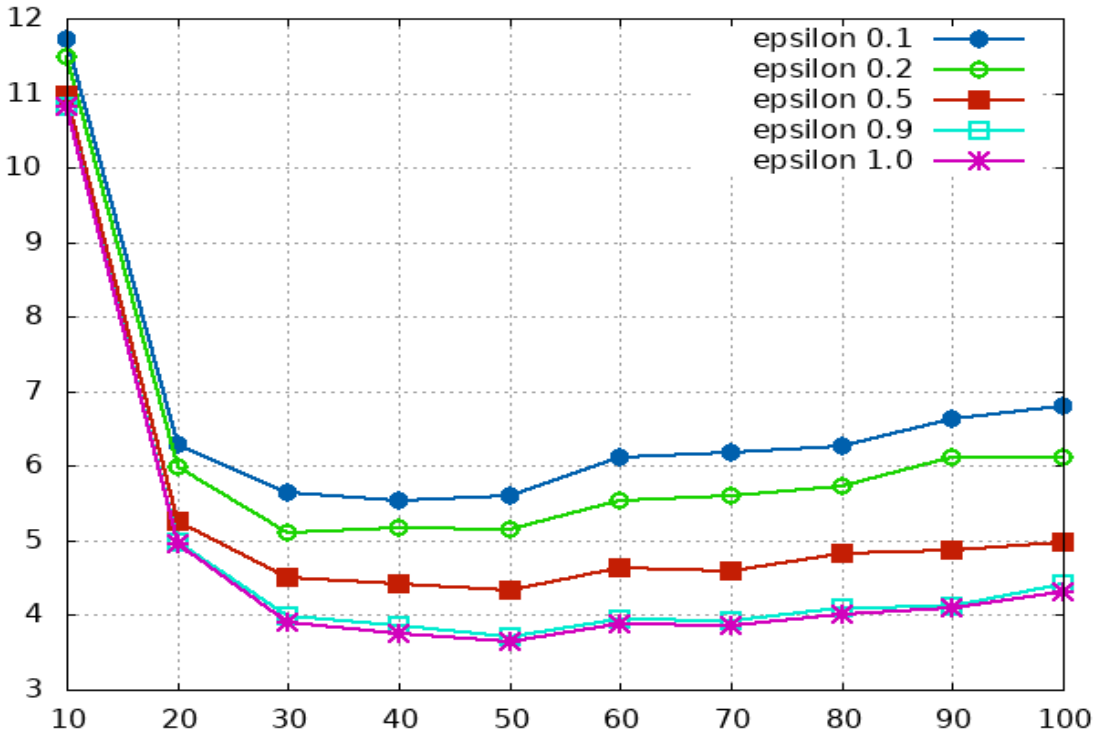


figura 2

Como se puede observar, cuanto mayor es el factor de suavizado epsilon, mejores resultados obtenemos, obteniendo los mejores resultados con epsilon = 1.
La reducción de las dimensiones de los datos usando PCA también hace que mejore la clasificación cuando mas aumenta hasta llegar a la reducción a la dimensión 50, a partir de la cual empieza a empeorar en todos los casos de epsilon, en dimensión PCA 50 con epsilon = 1 obtenemos la mejor tasa de error, siendo esta de 3,64%.
En comparativa, el clasificador multinomial con epsilon = 1 que se nos da de ejemplo en el boletín, tiene un error de clasificación de 14.28%.

Ejercicio 5.1

Implementa el paso M de estimación de los parámetros que gobiernan la mixtura de gaussianas y que se detallan en las Ecs. 10, 11 y 12.

En el fichero mixgaussian.m que se nos facilita, el cual lo renombraremos como Ejercicio5.m, se implementarán las ecuaciones que se encuentran en el boletín, en la parte del código donde aparece comentado HERE YOUR CODE FOR PARAMETER ESTIMATION. Tal que así:

```
% HERE YOUR CODE FOR PARAMETER ESTIMATION

auxzk = sum(zk);
  pkGc{ic} = auxzk/Nc;
  mu{ic} = Xc'*zk./auxzk;

  for k=1:K

    sigma{ic,k} = (zk(:,k).*(Xc - mu{ic}(:,k)'))'*(Xc - mu{ic} (:,k') /auxzk(k);
    sigma{ic,k} = alpha*sigma{ic,k}+(1-alpha)*eye(D);

  end
```

Ejercicio 5.2

Ejecutando el script Ejercicio5_2.m obtenemos el archivo ejercicio5_2.dat, con este archivo, ejecutando el script plot plot_5_2.plt obtenemos la siguiente gráfica (figura 3):

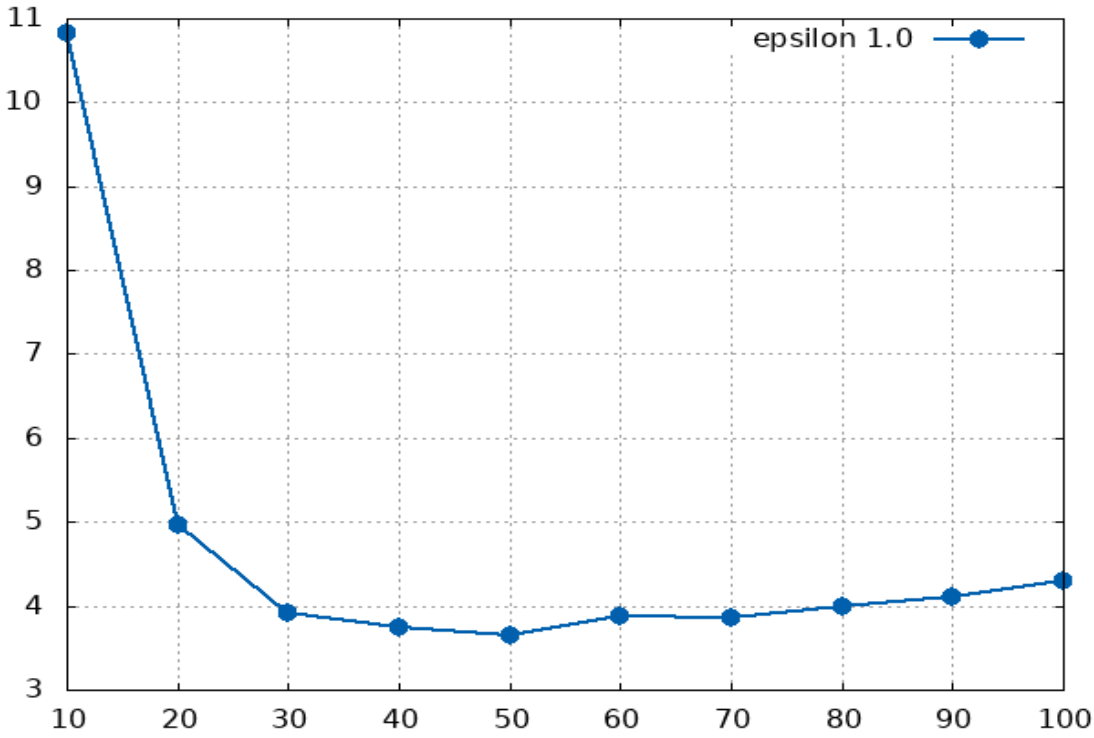


figura 3

Ejercicio 5.3

Realiza un experimento donde se evalúe el error de clasificación en función del número de componentes por mixtura para un número de componentes PCA y un valor α fijo. Representa gráficamente los resultados obtenidos en forma de una curva. Genera curvas de resultados adicionales utilizando diferente número de componentes PCA. Compara los resultados obtenidos con los reportados en la tarea MNIST.

Tal y como se nos indica en el boletín, haremos un experimento con diferentes componentes por mixtura y un factor de suavizado de 0.9, ejecutando nuestro script de octave (Ejercicio5_3.m) y el correspondiente script plot (plot_5_3.plt), obtenemos la siguiente gráfica (figura 4):

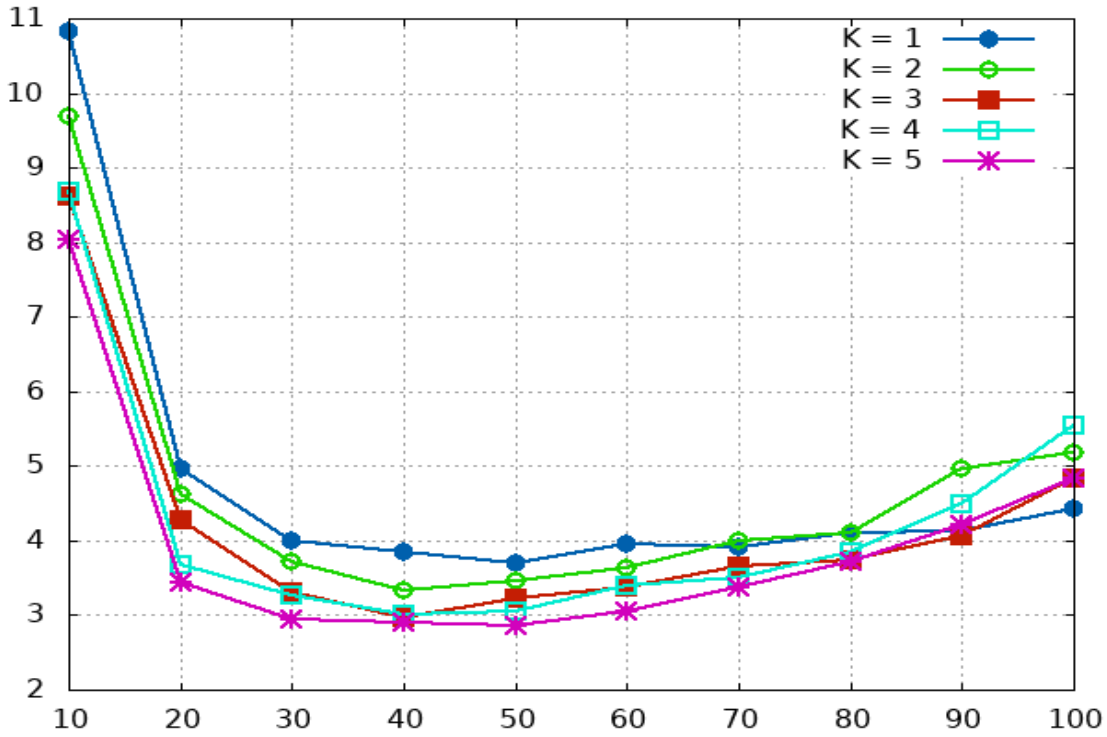
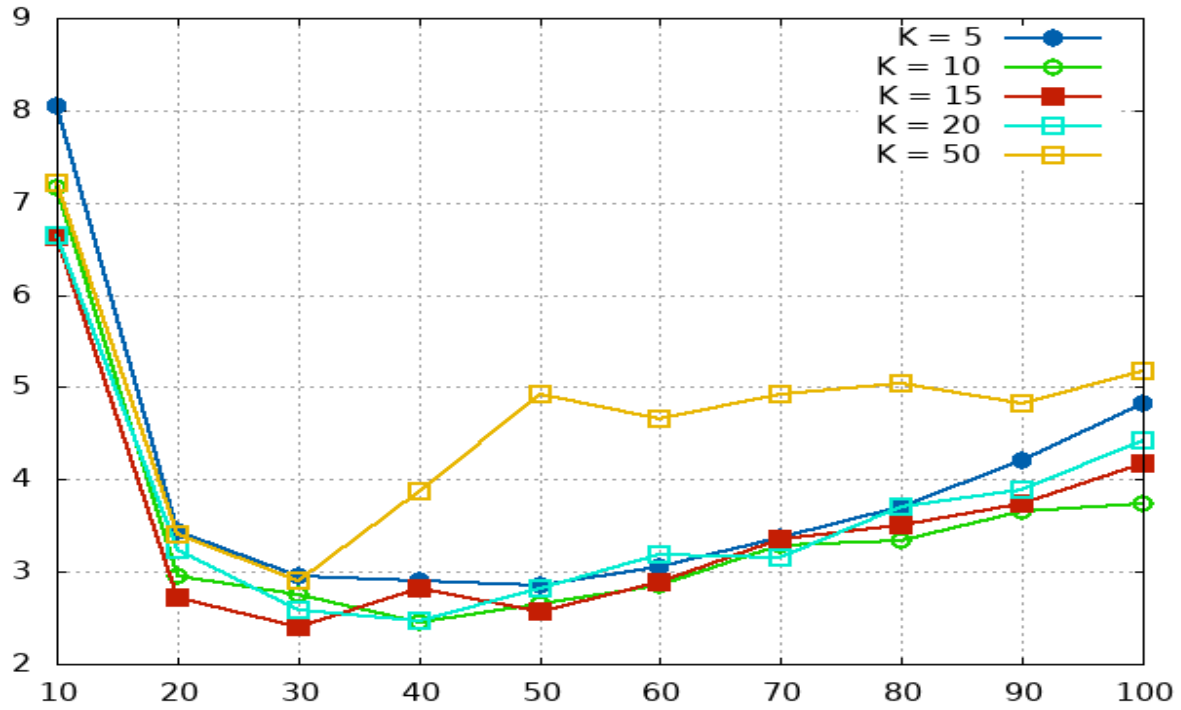


figura 4

Como se puede ver, se comporta igual que en el ejercicio 4.2, el error de clasificación con reducción de la dimensionalidad PCA disminuye hasta la dimensión 50, donde a partir de ella empieza a empeorar. También se observa que a cuantas más mixturas, mejor es el resultado, por tanto, podemos observar que a dimensión 50 y 5 mixturas, tenemos el mejor resultado, que es una tasa de error de 2.85%.

Visto que el error cada vez es menor cuantas más mixturas usamos, vamos a hacer un experimento con un número de mixturas mucho mayor que 5, en concreto con un K = 10, 15, 20 y 50.
La información de dicha gráfica se encuentra en el archivo ejercicio5_3_ .dat, donde con el script plot_5_3_2.plt obtenemos la siguiente gráfica (llamada ej5_3_2.png):



Como podemos observar, a cuanta mayor K, no siempre mejora el error, con K = 5, 10, 15 y 20, se ve claro que no siempre a más mixturas mejores resultados, por ejemplo, en reducción PCA 40, el error de clasificación con 15 mixturas es más alto que el error con 10 mixturas, también se aprecia en reducción PCA 60, dónde K = 20 es peor que todas sus anteriores (5,10 y 15).
A la vista de dichas pruebas, las cuales nos muestran que a cuantas más mixturas no siempre es mejor el error de clasificación, se ha decidido probar con un número muy alto de mixturas (K = 50), donde a partir de la reducción PCA 30, el error de clasificación se dispara en comparativa con las demás pruebas, por lo que podemos afirmar que a más mixturas, no se mejora el error.
En este experimento, la menor tasa de error conseguida ha sido de 2.4% y se ha hallado con 15 mixturas y reducción de dimensionalidad 30, siendo el factor de suavizado de 0.9 como en todo el ejercicio 5.3.