Anno Accademico: 2017/2018 Studente: Martini Davide

Sistemi Intelligenti Lab Homework

Classify SPAM/NON SPAM e-mails using SVM

L'obiettivo dell'esperienza è poter distinguere, in base alla frequenza di particolari parole e caratteri, se una mail può essere classificata o meno come spam. Questo è stato possibile tramite l'allenamento e la fase di test successivamente, di una soft margin Support Vector Machine (SVM). Il dataset è stato suddiviso in modo tale da avere il 50% dei dati a disposizione per la fase di training, e l'altro 50% per la fase di test. È stato fornito un dataset con dati grezzi, sia con dati in forma standard, in modo da poter confrontare l'efficienza e l'importanza de dati di partenza per ottenere un risultato migliore. Di seguito verranno riportate le risposte ai quesiti posti.

1. Comment the structure of the first sample in the **spam_train_std.txt**

```
1 1:-0.366938 2:-0.167405 3:9.43276 4:-0.0557516 5:2.41953 6:-0.366815
7:-0.286238 8:-0.295655 9:-0.354939 10:-0.394713 11:-0.314176 12:0.529747
13:-0.306858 14:-0.208795 15:-0.191982 16:-0.309682 17:-0.334178
18:-0.353622 19:-0.3703 20:-0.191408 21:-0.688461 22:-0.118089 23:-0.310492
24:-0.262914 25:-0.322417 26:-0.288082 27:0.0675372 28:-0.225349
29:-0.165064 30:-0.235041 31:-0.168662 32:-0.147569 33:-0.18829 34:-0.15032
35:-0.232195 36:-0.252764 37:-0.336529 38:-0.0529935 39:-0.200771
40:-0.176668 41:-0.110458 42:-0.169817 43:-0.210499 44:-0.140028
45:-0.306683 46:-0.189508 47:-0.0826673 48:-0.0999146 49:-0.163659
50:3.0189 51:-0.133191 52:-0.439094 53:-0.317204 54:-0.0926682
55:-0.111834 56:-0.203958 57:-0.403312
```

   Il contenuto dell'immagine riportata sopra è l'esempio di una mail standardizzata. Questa contiene:
   - Il primo numero denota se la mail viene considerata spam (1) oppure no-spam (0).
   - Il resto degli attributi indicano una particolare parola o carattere e quanto frequentemente appaiono nel testo della mail. Gli attributi run-length (55-57) misurano la lunghezza delle sequenze di lettere maiuscole consecutive.

2. Find the best tuning of the parameters for the three kernels using the 10-fold cross validation starting from the standardized training set with the grid search method:
   - **Linear** (C)
   - **Radial Basis Function (RBF)** (C, γ)
   - **Polynomial** (C, γ, α, d).

   Per il kernel di tipo lineare si è proceduto con il comando:
   ./svm-train –s 0 –t 0 –c "x" –v 10 ./svm_dataset/svm/spam_train_std.txt
   dove al posto di "x" sono state testati valori differenti per il costo. I valori presi in considerazione sono state tutte le potenze di 2, da 2^{-15} a 2^{15}.
   Alla fine di tutti i test il valore per la variabile C migliore è stato 0.039062, che porta ad una previsione di accuratezza nella fase di training del 92.6%.
   Per il kernel di tipo polinomiale si è proceduto con il comando:
   ./svm-train –s 0 –t 1 –c "x" –g "y" –d "k" –r "z" –v 10 ./svm_dataset/svm/spam_train_std.txt
   dove al posto di "x" sono state testati valori differenti per il costo, al posto di "y" i valori per la variabile γ, al posto di "k" i vari valori del grado del polinomio e al posto di "z" i valori del

possibile coefficiente di grado 0 presente nel polinomio. I valori presi in considerazione sono state tutte le potenze di 2, da 2^{-15} a 2^{15}, sia per la variabile C che per la variabile γ. Per la variabile d invece sono stati considerati i valori tra 2 e 5 compresi con passo 1. Non si è deciso di partire da 1 in quanto la funzione sarebbe risultata lineare. Per α invece sono stati considerati i valori da -5 a 5 compresi di passo 1.
Alla fine di tutti i test il valore per la variabile C migliore è stato 512, per γ 0.000015, per d 5 e per α il valore migliore è stato 1. Questi valori hanno portato ad una previsione di accuratezza nella fase di training del 92.8%.
Per il kernel di tipo radial basis function si è proceduto con il comando:
./svm-train –s 0 –t 2 –c "x" –g "y" –v 10 ./svm_dataset/svm/spam_train_std.txt
dove al posto di "x" sono state testati valori differenti per il costo, al posto di "y" i valori per la variabile γ. I valori presi in considerazione sono state tutte le potenze di 2, da 2^{-15} a 2^{15}, sia per la variabile C che per la variabile γ. Alla fine di tutti i test il valore per la variabile C migliore è stato 1.9 e per γ 0.007422. Questi valori hanno portato ad una previsione di accuratezza nella fase di training del 92.6%.

3. Assess the cross validation error with the increasing the complexity of the chosen model. When do you think it should be appropriate to stop the training phase in terms of complexity? Why?

   L'interruzione della fase di training credo debba avvenire quando l'errore supera del 10% il valore minimo di errore riscontrato fino a quel momento. Nei tre casi analizzati quindi si possono ammettere costanti fino a circa il 20% di errore. Non ha senso continuare a scendere con l'accuratezza del modello in quanto sarebbe poco utile la sua applicazione su dati mai analizzati, in quanto la percentuale di successo si abbassa.

4. Once you choose the best kernel and the corresponding tuning of parameters, predict the output for the standardized test set, using the learned model and assess the classification accuracy. For the sake of completeness evaluate the performance also for the other two kernels using their best configurations of parameters. Are the results consistent (in terms of accuracy) with the previous ones, obtained during the train phase?

   I vari kernel con le loro variabili:

	C	γ	α	d	Accuracy training	Accuracy test
Linear	0.039062				92.6%	92.6%
Polynomial	512	0.000015	1	5	92.8%	92.8%
Radial Basis Function	1.9	0.007422			92.6%	92.6%

   Il kernel quindi con accuratezza migliore risulta essere quello polinomiale, anche se di poco. I risultati di test risultano quindi coerenti con quelli della fase di training essendo la percentuale di accuracy trovata, identica. Questo valore risulta positivo, in quanto se la percentuale fosse risultata inferiore nella fase di test rispetto a quella di training si sarebbe presentata una situazione di overfitting, cioè ci si sarebbe concentrati troppo sulle caratteristiche dei dati di training piuttosto che sull'allenamento del sistema.

5. Train and test using a soft margin SVM for each kernel with the best three configurations found in the previous steps, using not standardized data. Note and comment the differences in

terms of computational time and classification accuracy. Do the three kernels perform in the same way?

	Accuracy training	Accuracy test
Linear	88.0%	89.2%
Polynomial	85.4%	89.8%
Radial Basis Function	77.0%	74.8%

Si nota immediatamente che con il dataset di dati non standardizzati si ottengono risultati solamente peggiori del 3% per i kernel lineare e polinomiale, il kernel radial basis function cala di quasi 20% e quindi questa funzione è quella più sensibile alla qualità dei dati. Rimane ancora il polinomiale il kernel con l'accuratezza migliore che si nota solo dopo la fase di test, in quanto nella fase di training il lineare aveva la percentuale di successo più elevata. Una previsione troppo ottimistica era stata fatta nella fase di training per il kernel radial basis function che vede scendere la sua accuratezza nella fase di test rispetto a quella di training.