4
Support Vector Machines

Support Vector Machines (SVM) é um algoritmo de aprendizado de máquina para classificação binária. Desenvolvido por Vapnik et al. em 1992 (19), tem sido bastante utilizado para a tarefa de classificação de sentimento.

O SVM trabalha apenas com dados numéricos. Sua entrada é um conjunto de vetores, cada um representando um exemplo. O conjunto de treino contém os exemplos já separados nas duas classes: positivo ou negativo. Cada vetor é mapeado para o espaço, e durante o treinamento o SVM busca o hiperplano com a máxima distância Euclidiana entre essas duas classes. Em seguida, o SVM recebe o conjunto de teste, também em formato de vetores, com os itens a serem classificados. Cada item é mapeado para o espaço e dependendo de sua posição relativa ao hiperplano separador, ele é classificado em uma das duas categorias.

Como o algoritmo de aprendizado não consegue processar textos, precisamos de uma camada adicional de representação. Para transformar o texto em dados numéricos é criado um dicionário de palavras com todas as palavras existentes em todos os documentos. Cada palavra é associada a uma coordenada do vetor. Em cada coordenada do vetor, é atribuído o valor 1 se aquela palavra existe naquele documento e zero caso contrário. Com esses vetores-linha, geramos uma matriz grande, porém esparsa, onde o número total de

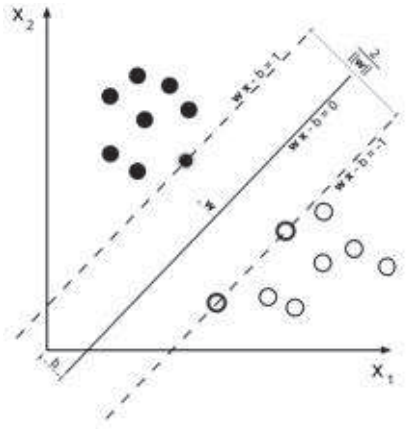

Figura 4.1: Hiperplano com a Máxima Distância

palavras é geralmente muito superior ao número de palavras em cada documento.

O classificador que utilizamos é o LibLINEAR (02), uma versão aprimorada do LibSVM (03), consideravelmente mais rápido que o último. Podemos trabalhar com um núcleo linear pois o tamanho dos vetores é muito maior do que o número de vetores.

A seguir descrevemos um exemplo prático para demonstrar como é feita a transformação de texto em dados numéricos.

Suponha que nossos documentos se resumem aos dois exemplos a seguir:

Documento 1 - Eu não gostei do filme.
Documento 2 - Gostei dos atores e do filme.

A seguir mostramos como ficariam os vetores. A primeira linha mostra o dicionário de palavras, as duas linhas seguintes mostram para cada documento como ficariam os vetores. Nas coordenadas em que a palavras existe no documento é preenchida com "1"enquanto que nas posições em que a palavra não existe é preenchida com "0".

Dicionário	Eu	não	gostei	do	filme	dos	atores	e	.
Documento 1	1	1	1	1	1	0	0	0	1
Documento 2	0	0	1	1	1	1	1	1	1

Tabela 4.1: Vetores de entrada para o SVM

Essa é uma representação por presença, 1 se a palavra está presente e zero caso contrário. Porém foi feito um experimento utilizando uma representação por frequência, onde no lugar do 1 coloca-se o número de vezes que a palavra aparece no exemplo. Esse experimento é menos eficiente que o modelo por presença, por isso utilizamos a representação presencial nesse trabalho.

Quando é utilizada a representação por frequência ou quando temos outros atributos que não sejam apenas o dicionário de palavras, por exemplo os atributos estruturais que será mostrado mais adiante, é necessário fazer um escalonamento para o mesmo intervalo. Isso evita que atributos em intervalos grandes tenham peso maior que atributos em intervalos pequenos. Todos os atributos desse trabalho são escalonados para o intervalo [0,1].