Minimização de Erro em Estimador de Densidade Kernel Por Meio da Projeção no Espaço das Verossimilhanças - Introdução a Reconhecimento de Padrões

Murilo Vale Ferreira Menezes - 2013030996

Novembro de 2016

1 Kernel Density Estimator

O KDE (Kernel Density Estimator - Estimador de Densidade Kernel) é uma técnica de estimar a densidade de uma dada distribuição, muito útil para classificadores Bayesianos. De acordo com um certo conjunto de pontos de treinamento, o KDE estima uma função densidade de probabilidade para cada ponto, somando-as para obter-se a densidade total do conjunto. A função para um dado ponto é denominada Kernel, e é dada pela seguinte expressão:

$$K(\frac{x-xi}{h}) = \frac{1}{2.\pi}.e^{-\frac{1}{2}\cdot(\frac{x-xi}{h})^2} \quad (1)$$

Dados os Kernels, podemos obter a estimativa de densidade simplesmente somando-os com um fator de normalização. Para um dado ponto x e N pontos de treinamento, temos então a seguinte probabilidade estimada:

$$p(x) = \frac{1}{N}.\sum_{i=1}^{N}\frac{1}{h}.K(\frac{x-xi}{h}) \quad (2)$$

Vemos então que podemos estimar a densidade tendo em mãos somente os dados de treinamento e o parâmetro h, que determina a forma do Kernel em cada ponto de treinamento, tendo a função análoga à do desvio padrão em uma função densidade de probabilidade normal. Escolher um bom valor para h é um passo importantíssimo no projeto do estimador. Abaixo podemos ver o comportamento da estimativa de densidadepara o benchmark *Spirals*.

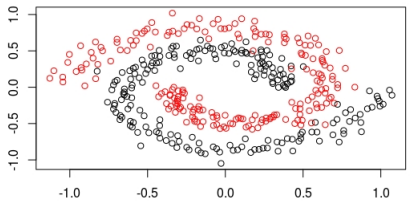

Figura 1: Benchmark Spirals

Variando-se h de 0 a 1, podemos obter o gráfico do erro para o conjunto de teste em função de h.

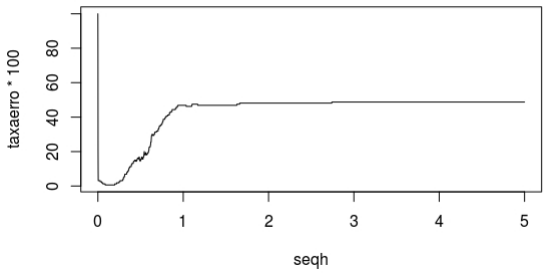

Figura 2: Erro de teste no conjunto Spirals

Podemos perceber que, para o conjunto de teste, h possui um valor ótimo, resultado em uma estimativa grosseira para um valor maior e tendendo ao *overfitting* para valores menores.

Neste trabalho, será explorada uma abordagem de encontrar este valor ótimo de h, baseado na análisa da projeção dos dados no espaço das verossimilhanças.

2 Projeção no espaço das verossimilhanças

A classificação Bayesiana funciona baseando-se na seguinte relação de probabilidades:

$$P(C|x) = \frac{P(x|C).P(C)}{P(x)} \qquad (3)$$

sendo C uma dada classe e x o ponto no qual se faz a classificação. A probabilidade $P(x|C)$ é denominada **verossimilhança**, e é fator determinante na classificação. Assim, para n classes, podemos mapear os pontos de acordo com as verossimilhanças para cada classe, em um espaço n-dimensional.

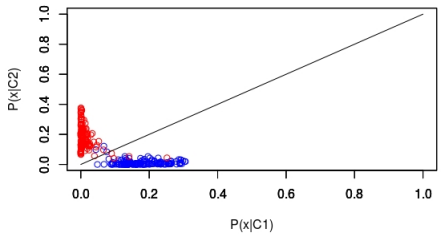

Figura 3: Mapeamento no espaço das verossimilhanças para um problema de classificação com duas classes

Neste trabalho, serão concebidas e aplicadas técnicas de análise deste espaço para diferentes valores de h, com o objetivo de encontrar uma relação do comportamento com um bom valor de h.

3 Descrição das técnicas utilizadas

Foram feitas oito técnicas para analisar o espaço das verossimilhanças, a saber:

- Distância euclidiana entre médias - Dos dados de cada classe, foi obtido o ponto de média, e, então, calculada a distância euclidiana entre eles. Por análise empírica do comportamento, foi decidido que esta distância deve ser minimizada.
- Distância de Mahalanobis - Foi calculada a distância de Mahalanobis, que leva em consideração a covariância, das classes entre si. Analisando o comportamento desta distância, queremos maximizá-la.
- Relação entre médias - Idealmente, queremos maximizar a verossimilhança de cada

classe em relação a seu eixo, bem como minimizá-la em relação ao eixo da classe oposta. Consideramos que a coordenada da média de uma classe i em relação ao eixo associado à classe j seja representada por med(Ci,Cj). Assim, queremos maximizar a seguinte relação: $\frac{med(C1,C1).med(C2,C2)}{med(C1,C2).med(C2,C1)}$.

- Relação entre desvios-padrão - Possui uma lógica parecida com a técnica anterior. Com base em observações da projeção, chegou-se à conclusão de diminuir o desvio-padrão de uma classe em relação ao eixo da classe oposta e aumentar em relação à sua própria. Chegamos então na seguinte relação, que se quer maximizar: $\frac{dp(C1,C1).dp(C2,C2)}{dp(C1,C2).dp(C2,C1)}$.

- Relação média-desvio-padrão - Esta técnica consiste na mistura das duas anteriores, analisando a seguinte relação: $\frac{med(C1,C1).med(C2,C2)}{dp(C1,C2).dp(C2,C1).med(C1,C2).med(C2,C1)}$.

- Análise dos desvios-padrão - Simplesmente uma análise dos desvios-padrão de uma das classes em relação a ambos os eixos. Em alguns casos, o máximo valor dos desvios-padrão coincide com valores de h próximos ao ótimo.

- Análise de covariância - Para as bases de dados em espaço bidimensional, foi avaliada a covariância entre as dimensões.

- Distância à reta - Considerando-se a reta de inclinação 1 no espaço das verossimilhanças como um separador ideal, é calculada a distância das médias a esta reta para cada valor na faixa de h. Em seguida, obtemos o mínimo valor na faixa.

4 Procedimentos

As técnicas anteriormente apresentadas foram testadas em seis bases de dados distintas de classificação binária, sendo três reais e três sintéticas. Todas as bases utilizadas se encontram na biblioteca *'mlbench'*. As bases foram:

- Spirals - Benchmark apresentado anteriormente, com dados de duas classes em forma de espiral.

- Circle - Benchmark que consiste em uma classe com dados distribuídos em círculo, cercados pelos dados de outra classe.

- 2D Normals - Base sintética que consiste em duas classes dispostas em distribuição normal, com superposição de algumas amostras.

- Winsconsin Breast Cancer - Base que classifica tipos de câncer de mama como benignos ou malignos, de acordo com algumas características.

- House Votes 84 - Base que contém votos de alguns congressistas dos EUA do ano de 1984, classificando-os em Republicanos ou Democratas.

- Pima Indian Diabetes - Contém amostras de alguns indicativos médicos, como glicose, insulina e massa corporal de índios da tribo Pima, e a informação de se o indivíduo possui ou não diabetes.

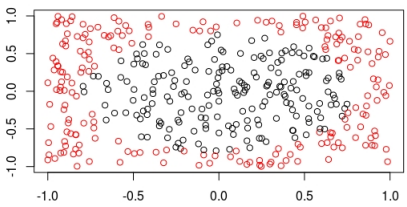

Figura 4: Base de dados Circle

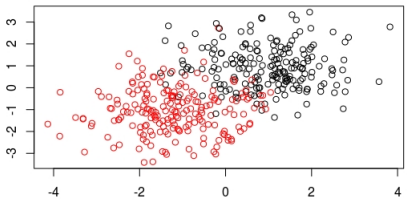

Figura 5: Base de dados 2D Normals

Primeiramente, a base de dados foi separada em conjuntos de teste e treinamento. Com base no conjunto de treinamento, foi calculado o erro para o conjunto de teste, dentro de uma faixa de valores de h. Destes valores, foi extraído o valor mínimo de erro obtido e o valor de h correspondente.

Em seguida, para a mesma faixa de h, os dados foram projetados no espaço das verossimilhanças. Foram projetados separadamente os dados de treinamento e os dados de teste. Então, foram obtidos os dados de posição das médias, desvios-padrão e covariâncias para a aplicação das técnicas acima.

Aplicadas as técnicas, salvou-se os resultados. O procedimento foi repetido por dez vezes, e no final se obteve a média dos erros mínimos para cada teste, bem como as diferenças dos erros obtidos pelas técnicas no espaço das verossimilhanças com o erro mínimo empiricamente obtido. Os resultados se encontram na tabela 1. A primeira linha é o valor médio do erro mínimo encontrado em cada teste. As outras, referentes a cada técnica, representam a diferença do valor de erro para o valor mínimo.

5 Análise

Analisando os resultados, podemos perceber primeiramente que o desempenho das técnicas varia muito do conjunto de treinamento para o conjunto de teste. Vemos, por exemplo, que a técnica da distância à reta de inclinação 1 teve resultados com um desvio muito pequeno ao ótimo para o conjunto de treinamento, enquanto teve um péssimo desempenho para as amostras de teste. Por outro lado, a distância de Mahalanobis teve uma ótima performance para o conjunto de teste, mas não foi tão bem com o conjunto de treinamento. Esta técnica, porém, encontrou bons valores de h para os conjuntos de treinamento nas bases reais Breast Cancer e House Votes. Isto ocorre pela larga faixa de valores de h que produzem erro satisfatório nestas bases, comparada à baixa faixa que produz um erro elevado. Este comportamento também ocorreu com a distância euclidiana entre médias. O comportamento típico para o método da distância euclidiana para o conjunto de treinamento do benchmark Spirals está a seguir.

Técnica	Spirals-Trn	Spirals-Tst	Circle-Trn	Circle-Tst	Norm-Trn	Norm-Tst	BCancer-Trn	BCancer-Tst	HVotes-Trn	HVotes-Tst	PIDiab-Trn	PIDiab-Tst
Mínimo	0,018	0,017	0,033	0,019	0,067	0,07	0,032	0,049	0,056	0,051	0,305	0,313
Mahalanobis	0,542	0,02	0,663	0,025	0,692	0,186	0,009	0,015	0,011	0,018	0,623	0,155
Dist. Médias	0,482	0,623	0,427	0,635	0,019	0,87	0,015	0,011	0,017	0,019	0,059	0,043
Rel. Médias	0,601	0,623	0,663	0,712	0,878	0,859	0,66	0,591	0,661	0,668	0,695	0,667
Rel. DP	0,113	0,113	0,12	0,281	0,35	0,182	0,155	0,251	0,444	0,079	0,355	0,587
Rel. Med. DP	0,56	0,32	0,442	0,568	0,819	0,604	0,461	0,43	0,494	0,255	0,681	0,633
DP C1-C1	0,012	0,011	0,108	0,102	0,013	0,067	0,66	0,647	0,661	0,668	0,07	0,055
DP C1-C2	0,236	0,296	0,042	0,044	0,013	0,01	0,03	0,044	0,06	0,058	0,07	0,055
Dist. Reta	0,016	0,623	0,012	0,635	0,019	0,87	0,011	0,011	0,017	0,019	0,04	0,687
Cov	0,207	0,234	0,406	0,348	0,066	0,549	0,011	0,21	0,017	0,019	0,07	0,055

Tabela 1: Resultados

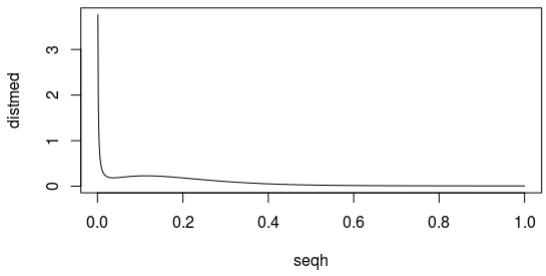

Figura 6: Variação da distância euclidiana entre médias com o parâmetro h para o conjunto de treinamento

A região de "joelho"do gráfico está intimamente ligada a um valor satisfatório de h. No gráfico obtido para a base Spirals, a região não é detectada como o mínimo, como podemos ver. Para outras bases, esta região representa o mínimo, o que acarreta na identificação de um bom valor de h.

As técnicas baseadas em relações entre médias e desvios-padrão com frequência produziram valores altos de erro, tanto para conjuntos de teste quanto para conjuntos de treinamento. Isto ocorre pela grande instabilidade da relação, já que os valores do denominador com frequência se aproximam muito do zero. Mesmo com uma análise detalhada, com grande resolução da faixa de h, estas técnicas não garantem uma performance mínima, o que as torna ruins para encontrar o valor ótimo. O mesmo ocorre com a técnica baseada na covariância.

As técnicas de análise de desvio-padrão mostraram desempenhos muito bons para algumas bases, porém não forneceram informação de interesse alguma em outras, o que não as tornam confiáveis.

A distância de Mahalanobis, para os conjuntos de teste, se comporta com a variação de h de acordo com o seguinte gráfico:

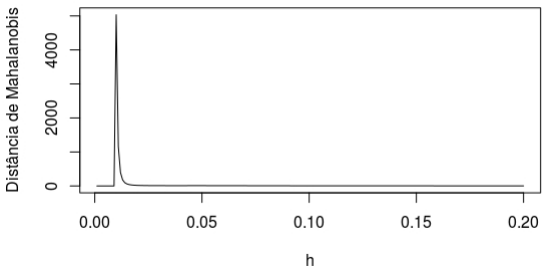

Figura 7: Comportamento da distância de Mahalanobis com h para o conjunto de teste

Vemos que, para uma faixa bem estreita de valores baixos de h, há um aumento repentino da distância, diminuindo logo em seguida. Nos pontos ao redor de onde este fenômeno acontece, o valor de erro é pequeno.

Para o treinamento, a distância tende a infinito quando h tende a zero, já que temos um caso de *overfitting* onde o valor das verossimilhanças é máximo para as classes próprias e mínimo para classes opostas. Desta forma, é difícil identificar um comportamento semelhante ao da figura 7, como mostra a figura 8:

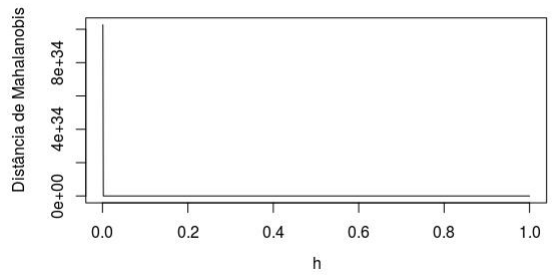

Figura 8: Comportamento da distância de Mahalanobis com h para o conjunto de treinamento

Porém, com uma análise detalhada - restringindo-se mais a faixa de h e diminuindo o passo de aumento- pode-se encontrar um aumento significativo na distância de Mahalanobis em uma faixa onde se encontra um h satisfatório. Assim, mesmo não retornando resultados adequados para o conjunto de treinamento, este método pode ser bem aproveitado para estas amostras, caso seja analisado de forma mais cuidadosa.

6 Conclusão

Neste trabalho, foi investigada uma alternativa para se analisar os efeitos do parâmetro h do método de estimativa de densidades KDE, baseada inteiramente no mapeamento para o espaço das verossimilhanças.

A análise foi executada de duas maneiras diferentes: com o mapeamento dos pontos de treinamento e dos pontos de teste. Com o mapeamento dos pontos de teste, alguns métodos são mais fáceis de analisar, como mostrado anteriormente, e o tempo de processamento é menor, uma vez que normalmente o tamanho do conjunto de teste é bem menor que o tamanho do conjunto de treinamento.

Porém, com a abordagem de mapeamento do conjunto de treinamento, tem-se um ganho valioso: apenas analisando-se o conjunto de treinamento, extrai-se dados necessários para um bom projeto de estimador, sem a necessidade de um conjunto de validação. Esta informação é intrínseca ao conjunto.

Com o resultado das análises deste trabalho combinado a uma análise mais aprimorada das técnicas propostas, pode-se então obter um bom valor para o parâmetro h do KDE, resultando em um bom classificador.