Vladimir G. Pestov

Métodos Matemáticos na Ciência de Dados: Introdução Relâmpago

Florianópolis, SC

2014

Vladimir G. Pestov

Métodos Matemáticos na Ciência de Dados: Introdução Relâmpago

Minicurso apresentado no III^o Colóquio de Matemática da Região Sul, realizado na Universidade Federal de Santa Catarina, em maio de 2014.

Florianópolis, SC

2014

Resumo

A ciência de dados, as vezes chamada de "a próxima grande coisa" (*"the next big thing"*), é um campo natural de pesquisa aplicada para os matemáticos. Em particular, a aprendizagem automática estatística é uma área de pesquisa fascinante, pelo menos a três níveis diferentes: como uma teoria matemática da grande profundidade e beleza, como uma direção do desenvolvimento de algoritmos, e como uma plataforma muito poderosa para aplicações práticas.

O ministrante do curso é um pesquisador em matemática pura, que está interessado também do desenvolvimento de novos algoritmos para análise de grandes conjuntos de dados. Em Novembro 2013, com uma equipe de 3 de seus estudantes de pós-graduação, ele ganhou o primeiro lugar na 4a Competição Internacional de Mineração de Dados de Segurança Cibernética (*4-th Cybersecurity Datamining Competition — CDMC'2013, Daegu, Korea, 3–7 do Novembro 2013*).

Este minicurso é uma introdução compacta e não tradicional aos métodos modernos de análise de grandes volumes de dados através da aprendizagem automática estatística, explicando a matemática para trás de alguns algoritmos que utilizou com sua equipe para vencer o evento.

Palavras-chaves: ciência de dados, aprendizagem automática estatística, classificador k-NN, consistência universal, aplicações borelianas, redução de dimensionalidade

Sumário

1 Problema de classificação binária

Vamos começar pela noção básica da *aprendizagem supervisonada*: o *problema de classificação binária*. Para tanto, tomamos uma experiência simples. Geramos $n = 1000$ pontos aleatórios no quadrado unitário $[0, 1]^2$, distribuidos uniformemente e independamentes um do outro. (A *distribuição uniforme* significa que a probabilidade de que um ponto x pertença a um pequeno quadrado $[a, a + \epsilon] \times [b, b + \epsilon]$ de lado $\epsilon > 0$ é proporcional (com efeito, igual) à área do quadrado, ϵ^2.)

Espera ver algo assim?

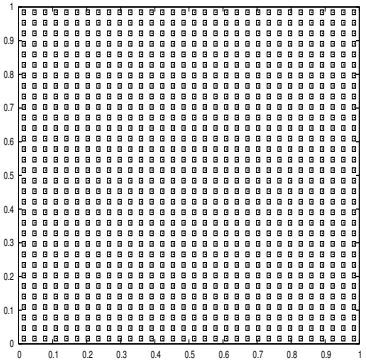

Figura 1 – Uma grade uniforme com $1024 = 32^2$ pontos.

Com efeito, isto não é o que o conjunto de dados resultante pode parecer. Em vez disso, veja figura 2.

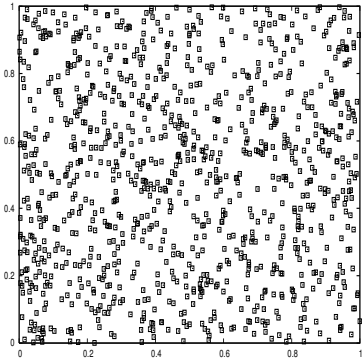

Figura 2 – Uma amostra aleatória de 1000 pontos tirados uniformemente do quadrado.

Note, em particular, a presença do que aparece como uma estrutura interna de dados significativa: os grandes buracos aqui e ali, agrupamentos de pontos... Estas são, na verdade, desvios aleatórios, não carregando nenhuma informação útil.

O nosso conjunto de dados,

$$X = \{x_1, x_2, \ldots, x_{1000}\},$$

é uma *amostra*. O quadrado $[0,1]^2$ é o *domínio.*

Agora dividimos os dados em duas classes: a classe A dos pontos sobre ou acima da diagonal (marcados pelos asteríscos) e a classe B dos pontos abaixo da diagonal (marcados pelos pequenos quadrados). Obtemos o que é chamado uma *amostra rotulada* (*labelled sample*). Ver Figura 3.

Observe um efeito visual interessante: parece que a fronteira entre as duas classes é uma curva ondulada, ao invés de

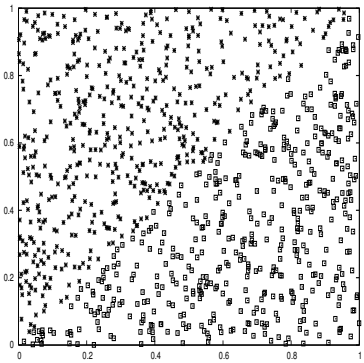

Figura 3 – Amostra rotulada.

uma linha reta! No entanto, você pode usar uma régua para convencer-se de que não há nenhum erro e os centros de todos os asteríscos estão realmente acima da diagonal, enquanto os centros de todos os quadrados estão abaixo.

Marcando os elementos de A com 1 e os elementos de B com 0, a nossa amostra rotulada pode ser escrita da seguinte maneira:

$$\sigma = (x_1, x_2, \ldots, x_{1000}, \varepsilon_1, \epsilon_2 \ \ldots, \varepsilon_{1000}),$$

onde por valor ε_i do rótulo do ponto x_i, temos $\varepsilon_i \in \{0, 1\}$, $i = 1, 2, \ldots, 1000$. (Certo, ao lado dos rótulos 0 e 1 pode se usar, por exemplo, -1 e $+1$...)

Neste exemplo "de brinquedo" a dimensão dos dados é 2, e o conjunto de dados pode ser visualizado, o que ajuda muito para determinar a sua estrutura. Cada ponto é um elemento de $\mathbb{R}^2$, representado por duas coordenadas, $x_i = (x_i^{(1)}, x_i^{(2)})$. A amostra rotulada σ pode ser tratada como um subconjunto (or-

denado) de $[0,1] \times [0,1] \times \{0,1\}$, e escrita como uma matriz de dimensão $1,000 \times 3$: cada linha $(x_i^{(1)}, x_i^{(2)}, \varepsilon_i)$ representa um elemento de X, assim que seu rótulo. Essa representação matricial dos conjuntos de dados é bastante comum. De uma maneira mais abstrata, podemos escrever

$$\sigma \in ([0,1] \times [0,1])^n \times \{0,1\}^n.$$

Chegamos ao sequinte *problema de classificação binária*: a partir da amostra rotulada σ, construir uma função

$$T\colon [0,1]^2 \to \{0,1\}$$

(chamada *classificador,* ou *preditor,* ou *função de transferência*), definida sobre todo o domínio, que seja capaz de predizer com confiança um rótulo não só para os dados existentes, mas também para novos dados. Podemos dizer que esse é o problema central da aprendizagem automática estatística *supervisionada.*

Claro que sabemos a resposta para nosso "problema de brinquedo": ela é dada pelo classificador de verdade

$$T_{true}(x) = \eta(x^{(1)} - x^{(2)}),$$

onde η é a *função de Heaviside,*

$$\eta(x) = \begin{cases} 1, & \text{se } x \geq 0, \\ 0, & \text{se } x < 0. \end{cases}$$

Mas se o problema for mostrado a alguma outra pessoa (ou máquina), que não sabe como as duas classes A e B foram formadas, você pode obter outras respostas. Por exemplo, o seu próprio córtex visual, ao analisar a imagem na figura 3, sugere separar as duas classes com uma linha ondulada! Um tal

classificador poderia não ser exato, mas estar perto da verdade para ser aceitável. As chances de *classificação errônea* (o *erro de classifição*) para um novo ponto de dados seriam relavamente pequenos.

Alguém pode sugerir a seguinte solução simplista: atribuir o valor 1 a todos os pontos de dados atuais que estão acima da diagonal, e o valor 0 a todos os outros pontos, atuais e futuros:

$$T(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

Este classificador dá uma resposta correta para todos os pontos atuais $x_i \in X$, $i = 1, 2, \ldots, 1,000$. No entanto, se nós gerarmos aleatoriamente um novo ponto $y \in [0, 1]^2$, com probabilidade $1/2$ ele ficará acima da diagonal. Ao mesmo tempo, a probabilidade de escolher um ponto em X é zero. Assim, com probabilidade de $1/2$, o classificador T irá retornar um valor falso para y. Entre todos os pontos gerados no futuro,

$$x_{1001}, x_{1002}, \ldots, x_{1000+n}, \ldots,$$

aproximadamente metade deles serão classificados erroneamente. Para n suficientemente grande, o classificador T fornecerá uma resposta errada aproximadamente na metade dos casos — certamente um fracasso completo. Jogando a moeda equilibrada podemos conseguir a mesma taxa de sucesso de $1/2$, sem usar qualquer classificador, simplesmente atribuindo a um ponto um valor aleatório 0 ou 1.

Como podemos distinguir um bom classificador de um ruim? Ou seja, dado um classificador, T, existe uma maneira de verificar se T é susceptível de atribuir a *maioria* dos pontos de dados futuros à classe correta?

À primeira vista, o problema parece completamente intratável: como possivelmente podemos mostrar algo sobre os dados que ainda não existem? Na verdade, é quase incrível que – pelo menos dentro de um modelo teórico – tais predições podem ser feitas com um grau considerável de certeza.

Todavia, vamos deixar este problema para mais tarde. Consideremos um exemplo real: um conjunto de dados da competição CDMC'2013 de mineração de dados para o problema de deteção de intrusos numa rede, coletados por um sistema real IDS (Intrusion Detection System) na Coréia. (Para mais informações, consulte [24]. Este conjunto não está disponível publicamente, mas outros conjuntos semelhantes são, por exemplo [6]).

Cada linha da matriz corresponde a uma sessão, onde as 7 coordenadas são os valores dos parámetros da sessão. Existem $n = 77,959$ pontos de dados, incluindo $71,758$ sessões normais (sem intruso), rotuladas $+1$, e $6,201$ sessões ataque (com intruso), rotuladas -1. A Figure 4 mostra um extrato das 15 linhas da matriz.

O objetivo é de construir um classificador capaz de alertar de um intruso em tempo real com um erro mínimo e uma confiança alta.

O que seria o classificador mais natural de se usar, baseado em nossa experiência cotidiana e o senso comum?

Suponha que você queira vender o carro. Para determinar um preço razoável, você vai buscar algumas informações sobre a venda dos carros do mesmo modelo, idade, milhagem, até a cor. Em outras palavras, você busca um carro o mais semelhante ao seu, e a sua cotação de venda dá uma boa idéia do

```
                    ...........
+1 1:-1.00 2:-0.03 3:-0.09 4:-0.49 5:-0.05 6:-0.15 7:-1.08
+1 1:-1.00 2:-0.03 3:-0.09 4:-0.49 5:-0.05 6:-0.15 7:-1.08
+1 1:-1.00 2:-0.03 3:-0.09 4:-0.49 5:-0.05 6:-0.15 7:-1.08
+1 1:-1.00 2:-0.03 3:-0.09 4:-0.49 5:-0.05 6:-0.15 7:-1.08
+1 1:-0.67 2:-0.03 3:0.04  4:1.95  5:-0.05 6:-0.10 7:1.11
+1 1:-1.00 2:-0.03 3:-0.09 4:-0.49 5:-0.05 6:-0.15 7:-1.08
+1 1:-1.00 2:-0.03 3:-0.09 4:-0.49 5:-0.05 6:-0.15 7:-1.08
+1 1:-0.63 2:-0.03 3:0.03  4:1.89  5:-0.05 6:-0.10 7:1.11
+1 1:-0.59 2:-0.03 3:0.03  4:1.83  5:-0.05 6:-0.09 7:1.11
-1 1:-1.00 2:-0.03 3:-0.09 4:-0.49 5:-0.05 6:-0.15 7:-1.08
+1 1:-1.00 2:-0.03 3:-0.09 4:-0.49 5:-0.05 6:-0.15 7:-1.08
+1 1:-1.00 2:-0.03 3:-0.09 4:-0.49 5:-0.05 6:-0.15 7:-1.08
+1 1:-1.00 2:-0.03 3:-0.09 4:-0.49 5:-0.05 6:-0.15 7:-1.08
-1 1:-1.00 2:-0.03 3:-0.09 4:-0.49 5:-0.05 6:-0.15 7:-1.08
+1 1:1.09  2:-0.03 3:-0.02 4:-0.49 5:-0.05 6:-0.15 7:1.11
                    ...........
```

Figura 4 – Fragmente do conjunto de dados para deteção de intrusos na rede.

preço a escolher.

É exatamente como o *classificador de vizinhas mais próximas*, ou o *clasificador NN* (Nearest Neighbour Classifier) funciona. Dado um ponto qualquer y do domínio, Ω, buscamos o ponto x do conjunto de dados atual, X, mais próximo a y. O classificador NN atribui a y o mesmo rótulo que o rótulo de x. Obviamente, a fim de determinar o vizinho mais próximo, precisa de uma função de semelhança qualquer sobre o domínio:

$$S: \Omega \times \Omega \to \mathbb{R}.$$

Tipicamente, S é uma métrica, por exemplo, a métrica euclideana.

Voltando à venda do carro, provavelmente é mais ra-

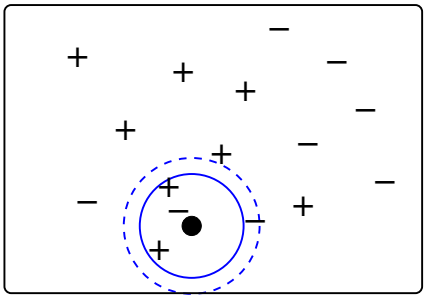

Figura 5 – O voto majoritário para $k = 3$, entre $+, +, -$, retorna $+$, a para $k = 4$, entre $+, +, -, -$, é indeciso.

zoável buscar mais de um carro semelhante ao seu, e determinar o preço baseado sobre uma variedade dos preços destes carros. Obtemos o *classificador de k vizinhos mais próximos*, ou *classificador k-NN*, onde k é um número fixo. Dada a amostra rotulada,

$$\sigma = (x_1, x_2, \ldots, x_n, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \in \Omega^n \times \{0, 1\}^n,$$

e o ponto da entrada $y \in \Omega$, o classificador k-NN escolhe k vizinhos mais próximos a y, $x_{i_1}, x_{i_2}, \ldots, x_{i_k} \in X$, e determina o rótulo de y pelo voto majoritário entre os rótulos $\varepsilon_{i_1}, \varepsilon_{i_2}, \ldots, \varepsilon_{i_k}$. Se a votação for indecisa (o que é possível se k for par), o rótulo de y é escolhido aleatoriamente.

Como podemos garantir que as previsões dadas por um classificador são confiáveis? Na prática, a técnica comumente usada é a *validação cruzada*. O conjunto de dados é dividido aleatoriamente no *conjunto de treinamento* (tipicamente, 75 a 90 por cento dos pontos) e o conjunto de avaliação (o restante 10 a 25 por cento). Somente os dados de treinamento são usados pelo algoritmo, e os dados de avaliação são aplicados para estimar o erro de predição.

Denotameros $T\colon \Omega \to \{0,1\}$ o classificador, X_t o conjunto de treinamento, e X_a o conjunto de avaliação:

$$X = X_t \cup X_a, \quad X_t \cap X_a = \emptyset.$$

O valor seguinte é o estimador estatístico do erro de predição (ou: erro de classificação) de T:

$$\frac{|\{i\colon x_i \in X_a,\ T(x_i) \neq \varepsilon_i\}|}{|X_a|}.$$

O procedimento é iterado muitas vezes, e o valor médio dos erros cada vez estimados serve como uma boa aproximação ao valor do erro verdadeiro de classificação do T.

Para aplicar classificadores aos conjuntos de dados concretos, é preciso escolher uma linguagem de programação. Teoricamente, qualquer linguagem pode ser utilizada: todas são equivalentes à máquina de Turing! Portanto, a linguagem utilizada mais comunamente em mineração de dados (até 2/3 dos casos, de acordo com algumas estimativas) é R [22], a lingagem de programação estatística, criada no Departamento de Estatística da Universidade de Auckland, Nova Zelândia e baseado em software livre (no formato do projeto GNU). Duas boas fontes introdutórias são [18] e [14]. A fonte mais abrangente com informações sobre R, *The R Book,* é disponível livremente na web [5].

Eu sugiro que você baixe a linguagem R seguindo as instruções em qualquer uma dessas fontes acima, e comece a experimentar com ela após os exercícios dos livros [18, 14].

Existem muitas implementações disponíveis do classificador k-NN em R, por exemplo, o classificador IBk do pacote RWeka, ou o do pacote FNN [9].

Exercício 1.1. *Baixar o conjunto de dados* Phoneme *[20], é treinar o classificador k-NN em R.*

Aplicando o classificador k-NN ao nosso conjunto de dados para deteção de intrusos na rede, obtemos um classificador cujo erro de classificação é ao torno de 0.3 %.

Certo, é um bom resultado. Todavia, se você participar numa competição, claro que todos outros participantes vão usar os classificadores padrão. Para melhorar o resultado, é preciso combinar as técnicas conhecidas com as novas abordagens. E antes de melhorar o desempenho do algoritmo, precisamos compreender o que pode ser melhorado, onde há um problema possivel?

Mas antes mesmo de examinar esta pergunta, temos uma ainda mais fundamental: por que nós esperamos que o classificador k-NN funcione, dê resultados confiáveis?

A única maneira de analisar as perguntas deste tipo é no formato de um modelo matemático da aprendizagem supervisionada. Este modelo é o tema da próxima aula.

2 Consistência universal

Os dados são modelados pelas *variáveis aleatórias,* o que é a noção básica da teoria de probabilidade. A fim de compreender esta noção, relembramos primeiramente a noção bem conhecida de *variável*, muito comum na matemática pura (geometria, álgebra, análise...) Eis alguns contextos típicos onde as variáveis fazem a sua aparição.

(1) Determinar os valores de x por quais

$$5x^2 - x + 3 = 0.$$

(2) Suponha que $t \in [0, 1]$. Então ....

(3) Sejam $x, y, z \in \mathbb{R}$ quaisquer. Suponha que $x < y$. Então $x + z > y + z$.

(4) Seja z um número complexo qualquer. O valor absoluto de z ....

Uma variável é um elemento qualquer (desconhecido) de um conjunto ($\mathbb{R}$ nos casos (1) e (3), $[0, 1]$ no (2), $\mathbb{C}$ no (4), etc.). As variáveis na teoria de probabilidade são de uma natureza ligeiramente diferente. Elas são denotadas habitualmente pelas letras *maiúsculas,* $X, Y, Z, \ldots$, a fim de distinguir das variáveis "usuais". Se X é uma variável aleatória (abreviamos: v.a.) real, isso significa duas coisas. Primeiramente, tudo como no caso de uma variável usual,

- X é un número real cujo valor exacto é desconhecido: $X \in \mathbb{R}$.

Mas tem mais das informações adicionais disponíveis. Mesmo se o valor de X é desconhecido, se sabe

- a probabilidade de X pertençer à cada região A de $\mathbb{R}$.

Em outras palavras, se $A \subseteq \mathbb{R}$ é uma parte de $\mathbb{R}$, então se conhece um número real entre 0 e 1 que fornece a probabilidade do evento $X \in A$. Este número é denotado por

$$P[X \in A],$$

e as informações conjuntas sobre os valores $P[X \in A]$ para todos A se chamam a *lei de probabilidade,* ou simplesmente a *lei* de X. Então, uma variável aleatória é uma variável "usual" munida de uma lei. Por exemplo, se $a, b \in \mathbb{R}$, $a \leq b$ são quaisquer, então se sabe a probabilidade

$$P[a < X < b]$$

de que o valor de X esteja entre a e b. A lei de uma variável aleatória se denota por uma letra grega, por exemplo, μ ou ν. É uma aplicação associando à cada região A de $\mathbb{R}$ um número real,

$$\mathbb{R} \supseteq A \mapsto \mu(A) = P[X \in A] \in [0, 1].$$

Eis alguns exemplos.

1. Uma variável aleatória de Bernoulli toma dois valores: 0 e 1, cada uma com a probabilidade 1/2:

$$P[X = 0] = \frac{1}{2} = P[X = 1].$$

Para calcular a lei de X, seja $A \subseteq \mathbb{R}$ um conjunto qualquer. Obviamente, se A contém ambos 0 et 1, então a probabilidade que $X \in A$ é igual à 1, é um evento certo. Se A não contém nem

0 nem 1, então o evento $X \in A$ é improvável, a sua probabilidade é 0. Afinal, se A contém exatamente um dos pontos $\{0, 1\}$, então a probabilidade do evento $X \in A$ é $1/2$:

$$P[X \in A] = \begin{cases} 1, & \text{se } 0, 1 \in A, \\ \frac{1}{2}, & \text{se } 0 \in A \text{ e } 1 \notin A, \\ \frac{1}{2}, & \text{se } 0 \notin A \text{ e } 1 \in A, \\ 0, & \text{se } 0 \notin A,\ 1 \notin A. \end{cases}$$

Uma variável de Bernoulli modela uma jogada única de uma moeda justa, onde a probabilidade de dar "coroa" (o valor 1) é $1/2$, a mesma que a probabilidade de dar "cara" (o valor 0).

De maneira mais geral, se a moeda não é justa, então a probabilidade de dar "coroa" pode ser um valor qualquer $p \in [0, 1]$,

$$P[X = 1] = p,$$

é a probabilidade de dar "cara" é

$$P[X = 0] = 1 - p = q.$$

A lei de probabilidade de uma variável aleatória real, X, é completamente determinada pela sua *função de distribuição*, Φ. É uma função real dada por

$$\Phi(t) = P[X < t].$$

É facil de calcular a função de distribuição de uma v.a. de Bernoulli, veja Figura 6.

Se o conjunto dos valores da função de distribuição de uma variável aleatória X é enumerável, então X é dita *discreta.* Por exemplo, a variável aleatória de Bernoulli é discreta.

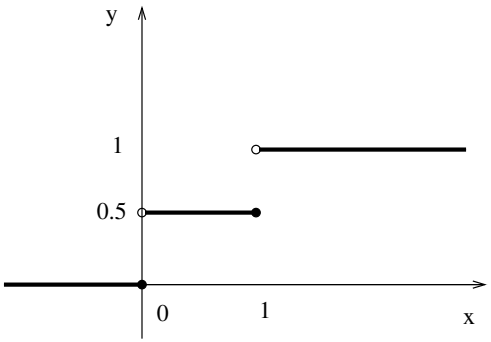

Figura 6 – Função de distribuição de uma variável aleatória de Bernoulli.

2. Uma variável aleatória de lei *uniforme* com os valores no intervalo $[0, 1]$ é dada pela fórmula seguinte: quaisquer sejam $a, b \in \mathbb{R}$, $a < b$,

$$P[X \in (a, b)] = \int_a^b \chi_{[0,1]}(t)\, dt. \tag{2.1}$$

Aqui, $\chi_{[0,1]}$ nota a *função indicadora* do intervalo $[0, 1]$ (Figura 7):

$$\chi_{[0,1]}(t) = \begin{cases} 1, & \text{se } x \in [0, 1], \\ 0, & \text{caso contrário.} \end{cases}$$

Por exemplo,

$$P[0 \leq X \leq 1] = \int_0^1 \chi_{[0,1]}(t)\, dt = 1,$$

e

$$P\left[-\frac{1}{2} \leq X \leq \frac{1}{2}\right] = \int_{-\frac{1}{2}}^{\frac{1}{2}} \chi_{[0,1]}(t)\, dt = \frac{1}{2}.$$

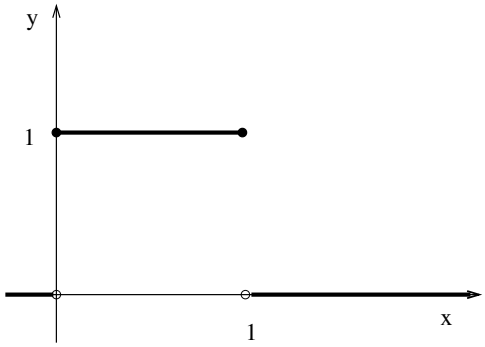

Figura 7 – Grafo da função indicadora do intervalo $[0, 1]$.

Se um intervalo (a, b) está contido em $[0, 1]$, então

$$\begin{aligned} P[X \in (a,b)] &= \int_a^b \chi_{[0,1]}(t)\, dt \\ &= \int_a^b 1 \cdot dt \\ &= b - a. \end{aligned}$$

Em outras palavras, neste caso a probabilidade de que X pertença ao intervalo (a, b) é igual ao comprimento do intervalo.

Se a lei de uma variável aleatória é dada pela integral, como na fórmula (2.1), então a função sobre integral é dita a *densidade* de X. A densidade de uma v.a. uniforme é a função indicadora:

$$\chi_{[0,1]}(t).$$

Exercício 2.1. *Mostrar que a lei de Bernoulli não possui densidade.*

A função de distribuição de uma v.a. uniforme é calculada facilmente (Figura 8).

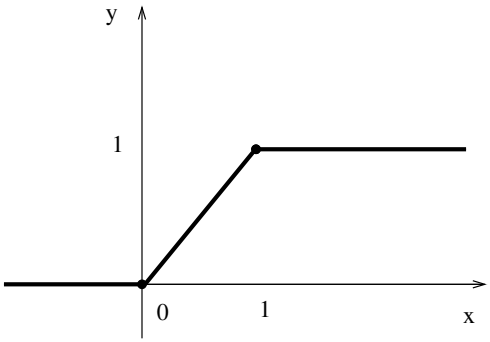

Figura 8 – A função de distribuição de uma variável aleatória uniforme.

Uma variável aleatória real X é dita *contínua* se os valores da sua função de distribuição preenchem o intervalo $[0, 1]$. A v.a. uniforme é obviamente contínua.

Exercício 2.2. *Seja X uma v.a. a qual possui densidade. Mostrar que X é contínua.*

Exercício 2.3 $(*)$**.** *Construir um exemplo de v.a. a qual não é contínua e não possui densidade.*

Exercício 2.4. *Construir um exemplo de v.a. nem discreta nem contínua.*

Uma variável aleatória real é *gaussiana* (ou: segue a lei *normal centrada e reduzida*, se X possui densidade dada por

$$\frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Em outras palavras, quais quer sejam $a, b \in \mathbb{R}$,

$$P[a < X < b] = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt.$$

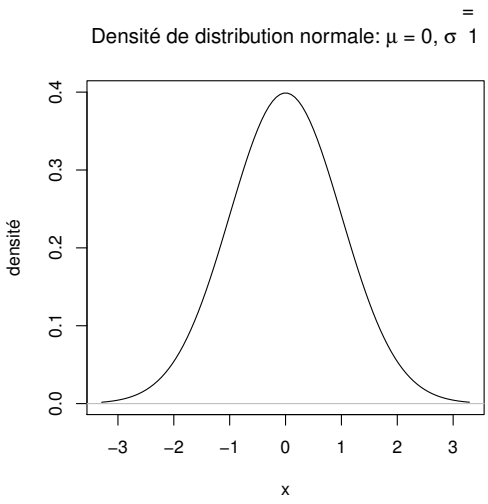

Figura 9 – A função de densidade da lei gaussiana.

A lei *semi-circular* é dada pela função de densidade

$$f(t) = \begin{cases} \frac{2}{\pi}\sqrt{1-t^2}, & \text{si } |t| \leq 1, \\ 0, & \text{se não.} \end{cases}.$$

Estritamente falando, o gráfo da densidade não é um semi-círculo, mas, melhor, uma semi-elipse – o fator normalizador $2/\pi \approx 0.637$ é necessário para que a probilidade de um evento certo seja igual a 1.

A noção de uma variável aleatória não é apenas o único conceito mais fundamental da teoria de probabilidade, mas é,

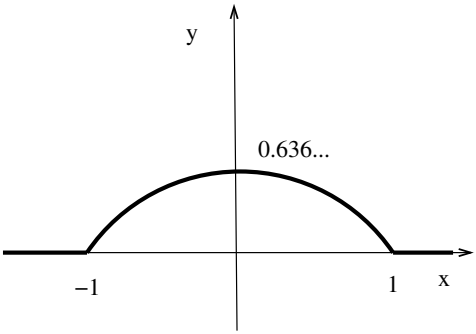

Figura 10 – A densidade da lei semi-circular.

sem dúvida, uma das mais importantes noções em todas ciências matemáticas. Alguns matemáticos argumentam que, eventualmente, os fundamentos da matemática devem ser alterados de modo que as variáveis aleatórias sejam tratadas juntamente com conjuntos...

Até agora, só vimos as variáveis aleatórias reais, com valores em $\mathbb{R}$. Mas elas podem assumir valores em domínios Ω mais gerais.

Seja Ω um domínio geral. Quais são as propriedades desejadas da lei, μ, de uma variável aleatória X com valores em Ω? Claro, os valores da lei pertencem no intervalo $[0, 1]$, e a probabilidade que $X \in \Omega$ deve ser 1:

(P1) $P[X \in \Omega] = \mu(\Omega) = 1$.

A probabilidade de x pertencer a união de uma família disjunta dos conjuntos A_i, $i \in I$ deve ser igual à soma das

probabilidades de que $x \in A_i$ para todos i:

$$Pr\left[X \in \bigcup A_i\right] = \sum Pr[X \in A_i].$$

Qual é o tamanho das famílias que devemos considerar? Se nós restringimos a propriedade às uniões finitas, a noção de probabilidade resultante é muito geral e fraca demais. Se, pelo contrário, permitimos as uniões de *todos* as famílias, a noção de probabilidade que obtemos é demasiado restritiva.

Exercício 2.5. *Seja X uma variável aleatória com valores em um conjunto Ω cuja lei possui a propriedade que, qualquer seja a família disjunta dos conjuntos $A_i \subseteq \Omega$, $i \in I$, $A_i \cap A_j = \emptyset$ por todos i, j, $i \neq j$, temos*

$$P\left[X \in \cup_{i\in I} A_i\right] = \sum_{i\in I} P[X \in A_i].$$

Mostrar que X é discreta.

A escolha mais natural e frutífera é a das famílias *enumeráveis*.

(P2) Se A_i, $i = 1, 2, 3, \ldots$ são disjuntos dois-a-dois, então $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

Como um corolário imediato, obtemos, no caso onde $A_1 = A$ e $A_2 = A^c = X \setminus A$:

(P2′) Se $A \subseteq \Omega$, então $P(A^c) = 1 - P(A)$.

Se μ é a lei de uma variável de Bernoulli (mais geralmente, de uma variável discreta), então o valor

$$\mu(A) = P[X \in A]$$

é bem definido qual seja um subconjunto $A \subseteq \Omega$ do domínio. Podemos esperar o mesmo para cada variável aleatória? A resposta é negativa. Com efeito, pode se mostrar que se uma v.a. X de lei μ não é discreta, então o valor $\mu(A)$ não pode ser definido por todos subconjuntos do domínio da maneira que as propriedades (P1) e (P2) sejam satisfeitas (assumindo o Axioma de Escolha).

Por esse motivo, somos forçados a restringir a coleção $\mathscr{B}$ dos subconjuntos $A \subseteq \Omega$, para as quais o valor $P[X \in A]$ é bem definido. O axioma (P1) implica que Ω sempre pertença à família $\mathscr{B}$. Segundo o axioma $(P2)$, se

$$A_1, A_2, \ldots \in \mathcal{B},$$

então a sua união pertença a $\mathscr{B}$ também:

$$\bigcup_i A_i \in \mathscr{B}.$$

Tendo em conta o axioma (P2′), concluímos que, se $A \in \mathscr{B}$, então $A^c \in \mathscr{B}$. Em breve, a família $\mathscr{B}$ deve contar Ω, os complementares de todos os seus membros, e as uniões de sub-famílias enumeráveis.

Se Ω é um espaço métrico, é razoável de exigir que a lei seja bem-definida para todas as bolas abertas:

$$B_r(x) = \{y \in \Omega \colon d(x, y) < r\}.$$

Isso é necessário, por exemplo, para conhecer a probabilidade do evento

$$[d(X, x) < r].$$

A *menor* família $\mathscr{B}$ que contém Ω, todas as bolas abertas, é fechada com relação aos complementares e uniões de sub-famílias

enumeráveis, se chama a família de sub-conjuntos *borelianos* de Ω.

Um espaço métrico Ω é dito *separável* se existe um sub-conjunto enumerável A cujo fecho é Ω:

$$\bar{A} = \Omega.$$

Exercício 2.6. *Seja Ω um espaço métrico separável. Mostrar que cada subconjunto aberto e cada subconjunto fechado de Ω são borelianos.*

Exercício 2.7. *Mostrar os exemplos de subconjuntos borelianos de $[0,1]$ que não são nem abertos nem fechados.*

Uma função μ na classe $\mathscr{B}$ dos conjuntos borelianos de Ω com valores em $[0,1]$ que satisfaz (P1) e (P2) é uma *medida de probabilidade boreliana*. Cada medida de probabilidade sobre Ω é a lei de uma variável aleatória com valores em Ω.

Sejam Ω e W dois espaços métricos, e $f\colon \Omega \to W$ uma função. Seja X uma variável aleatória com valores em Ω. Então $f(X)$ é uma variável aleatória com valores em W. A lei, ν, de $f(X)$ é a *imagem direita* da lei μ de X pela f: se $B \subseteq W$, então

$$\nu(B) = \mu(f^{-1}(B)).$$

Demonstração:

$$P[f(X) \in B) = P[X \in f^{-1}(B)].$$

A lei ν é as vezes denotada

$$\nu = f_*(\mu).$$

A única condição necessária sobre f é que a imagem inversa de cada sub-conjunto boreliano $B \subseteq W$ por f seja boreliano. Uma tal função se chama *função boreliana*. Pode se verificar que $f: \Omega \to W$ é boreliana se e somente se a imagem inversa de cada sub-conjunto *aberto* de W é boreliana. Em particular, cada função contínua é boreliana, mas as funções borelianas são muito mais numerosas.

Exercício 2.8. *Construir uma função boreliana discontínua.*

Se temos mais de uma variável aleatória,

$$X_1, X_2, \ldots, X_n, \ldots,$$

tomando os valores, respectivamente, nos espaços $\Omega_1, \Omega_2, \ldots, \Omega_n, \ldots$, então elas podem ser combinados numa única variável aleatória, tomando os valores no produto dos espaços Ω_i:

$$X = (X_1, X_2, \ldots, X_n, \ldots) \in \Omega_1 \times \Omega_2 \times \ldots \times \Omega_n.$$

A lei μ da variável X é chamada a *lei conjunto* das variáveis $X_1, X_2, \ldots, X_n, \ldots$. Notação:

$$\mu = \otimes_{i=1}^{\infty} \mu_i.$$

Esta μ é também chamada a medida produto das medidas de probabilidade $\mu_1, \mu_2, \ldots$.

As variáveis aleatórias $X_1, X_2, \ldots, X_n, \ldots$ são ditas *independentes* se, cada vez que A_i é um subconjunto boreliano de Ω_i, $i = 1, 2, \ldots$, temos

$$Pr[X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n, \ldots] =$$
$$Pr[X_1 \in A_1] \times Pr[X_2 \in A_2] \times \ldots \times Pr[X_n \in A_n] \times \ldots.$$

Por exemplo, sejam X e Y duas v.a., cada uma de lei uniforme sobre o intervalo $[0,1]$. Se X e Y são independentes, isso significa que a variável aleatória $Z = (X, Y)$ com valores no quadrado $[0,1]^2$ tem lei, μ, que é uniforme no quadrado: quaisquer sejam a, b, c, d, $a \leq b$, $c \leq d$, temos

$$\mu([a,b] \times [c,d]) = (b-a)(d-c).$$

Ao contrário, se, por exemplo, $Y = X$, então a lei da variável $Z = (X, Y)$ é concentrado na diagonal do quadrado: se $A_1, A_2 \subseteq [0,1]$ são disjuntos, então, obviamente,

$$P[X \in A_1, Y \in A_2] = 0,$$

de onde é fácil de concluir que

$$\mu(\Delta) = 1,$$

onde

$$\Delta = \{(x,x) \colon x \in [0,1]\}.$$

Se $X_{1,2}, \ldots, X_n$ é uma sequência das variáveis aleatórias independentes distribuídas segundo a lei gaussiana em $\mathbb{R}$, então sua lei conjunto é a lei gaussiana n-dimensional em $\mathbb{R}^d$, determinada pela densidade

$$\frac{1}{(2\pi)^{n/2}} e^{-(t_1^2+t_2^2+\ldots+t_n^2)/2}.$$

Isso significa que, qualquer seja um sub-conjunto boreliano $A \subseteq \mathbb{R}^d$,

$$P[X \in A] = \frac{1}{(2\pi)^{n/2}} \int_A e^{-(t_1^2+t_2^2+\ldots+t_n^2)/2} dt_1 \ldots dt_n.$$

Se o espaço Ω é munido de uma estrutura linear (além da boreliana) — por exemplo $\mathbb{R}$, ou $\mathbb{R}^d$, ou espaço de Hilbert, etc. — então pode se definir a *esperança* de X:

$$\mathbb{E}X = \int_\Omega x \, d\mu(x).$$

A lei dos grandes números é o resultado mais básico de probabilidade. Seja $X_1, X_2, \ldots, X_n, \ldots$ uma sequência de variáveis aleatórias reais independentes identicamente distribuídas. Suponha que X_i são limitadas. Então, os valores médios de $X_1, X_2, \ldots, X_n$ convergem para a esperança comum de X_i *em probabilidade* quando $n \to \infty$:

$$\forall \epsilon > 0, \quad P\left[\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \mathbb{E}(X_1)\right| > \epsilon\right] \longrightarrow 0.$$

Por exemplo, se μ é a lei de Bernoulli com $P[X = 1] = p$, é $X_1, X_2, \ldots, X_n, \ldots$ é uma sequência das v.a. independentes que seguem a lei μ, então os valores da frequência de dar "coroa",

$$\frac{X_1 + X_2 + \ldots + X_n}{n},$$

concentram-se ao torno de $p = \mathbb{E}(X_i)$ quando $n \to \infty$.

Agora estamos prontos para descrever o modelo fundamental da aprendizagem automática estatística. O domínio Ω é um espaco métrico separável e completo (como, por exemplo, $\mathbb{R}^d$). Um ponto (x, ε) de dados rotulado, onde $x \in \Omega$, $\varepsilon \in \{0, 1\}$, é modelado por uma variável aleatória (X, Y) com valores no produto $\Omega \times \{0, 1\}$. Aqui, $X \in \Omega$ representa um ponto no domínio, e $Y \in \{0, 1\}$, o rótulo marcando o ponto. A lei conjunto de (X, Y) é uma medida de probabilidade, μ, sobre $\Omega \times \{0, 1\}$. Agora, o ponto $x \in \Omega$ é dito *instância* da v.a. X, e o rótulo ε é uma instância da v.a. Y.

Pode-se mostrar a existência de uma medida de probabilidade μ_Ω sobre Ω, e uma função boreliana $\eta\colon \Omega \to \{0,1\}$ (chamada a *função de regressão*), tais que a escolha de (X,Y) é efetuada como segue: $X \in \Omega$ é uma variável aleatória com a lei μ_Ω, e uma vez que a instância x de X é escolhida, o valor de Y é escolhido lançando a moeda com a probabilidade $\eta(x)$ de dar "coroa". (A maneira de formalizar tudo isso é através da probabilidade condicional...)

É importante de ressaltar que, mesmo se sempre suponhamos que a lei μ existe, ela é sempre desconhecida. Também, as vezes o mesmo ponto $x \in \Omega$ pode obter rótulos diferentes.

Um *classificador* é uma função boreliana

$$T\colon \Omega \to \{0,1\}.$$

Dado um classificador, o seu *erro de classificação* é o valor real

$$\begin{aligned} \mathrm{err}_\mu(T) &= P[T(X) \neq Y] \\ &= \mu\{(x,y) \in \Omega \times \{0,1\}\colon T(x) \neq y\}. \end{aligned}$$

O *erro de Bayes* é o ínfimo dos erros de classificação de todos os classificadores possiveis sobre Ω:

$$\ell^* = \ell^*(\mu) = \inf_T \mathrm{err}_\mu(T).$$

Pode-se mostrar que, com efeito, o ínfimo é o mínimo, atingido pelo *classificador de Bayes:*

$$T_{bayes}(x) = \begin{cases} 0, & \text{se } \eta(x) < \frac{1}{2}, \\ 1, & \text{se } \eta(x) \geq \frac{1}{2}, \end{cases}$$

$$\mathrm{err}_\mu(T_{bayes}) = \ell^*(\mu).$$

O significado do classificador de Bayes é puramente teórico, porque a função de regressão, η, é desconhecida, assim como a lei μ.

Uma *regra da aprendizagem* é uma aplicação associando a cada amostra rotulada, σ, um classificador, T. Dado uma amostra

$$\sigma = (x_1, x_2, \ldots, x_n, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n),$$

a regra produz um classificador, $T = \mathcal{L}_n(\sigma)$, que é uma função boreliana de Ω dentro $\{0, 1\}$.

De maneira mais formal, podemos dizer que uma regra de aprendizagem é uma família $\mathcal{L} = (\mathcal{L}_n)_{n=1}^{\infty}$, onde para cada $n = 0, 1, 2, \ldots$,

$$\mathcal{L}_n \colon \Omega^n \times \{0, 1\}^n \to \Omega^{\{0,1\}}.$$

As aplicações associadas de avaliação

$$\Omega^n \times \{0, 1\}^n \times \Omega \ni (\sigma, x) \mapsto \mathcal{L}_n(\sigma)(x) \in \{0, 1\}$$

devem ser borelianas.

Por exemplo, o classificador k-NN é uma regra de aprendizagem.

A amostra rotulada $(x_1, x_2, \ldots, x_n, \varepsilon_1, \ldots, \varepsilon_n)$ é modelada pela sequência $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ das variáveis independentes com valores em $\Omega \times \{0, 1\}$, seguindo a lei fixa porêm desconhecida, μ. Para cada n, a regra de aprendizagem só "vê" os n primeiros pares de variáveis.

A regra de aprendizagem $\mathcal{L}$ é chamada *consistente* se o erro de classificação converge para o erro de Bayes (o menor

possível) em probabilidade quando $n \to \infty$:

$$\forall \varepsilon > 0, \quad P\left[\mathrm{err}_\mu \mathcal{L}_n > \ell^*(\mu) + \varepsilon\right] \to 0 \text{ quando } n \to \infty.$$

Porque não conhecemos a lei subjacente, μ, precisamos que a regra de aprendizagem seja consistente para todas as leis possíveis. Isto leva à seguinte definição. A regra $\mathcal{L}$ é *universalmente consistente* se ela é consistente para cada medida de probabilidade μ sobre $\Omega \times \{0, 1\}$.

Teorema 2.9 (Stone [25])**.** *Suponha que $k = k_n \to \infty$ e $k_n/n \to 0$. Então o classificador k-NN em $\mathbb{R}^d$ (com a distância euclideana) é universalmente consistente.*

O teorema de Stone falha nos espaços métricos mais gerais, mesmo no espaço de Hilbert de dimensão infinita (cf. um exemplo em [4], páginas 351–352, baseado sobre a construção de Preiss [21]).

A prova original de Stone era bastante complexa. No entanto, vamos delinear a idéia vaga de uma prova alternativa [7], [4], baseada em um resultado importante de análise real.

Teorema 2.10 (Teorema de densidade de Lebesgue–Besicovitch)**.** *Seja μ uma medida de probabilidade boreliana sobre $\mathbb{R}^d$, e $f\colon \mathbb{R}^d \to [0, 1]$ uma função boreliana. Então, o valor médio de f numa vizinhança de μ-quase todo ponto x, converge para $f(x)$, quando o raio da vizinhança converge para 0:*

$$f(x) = \lim_{\epsilon \downarrow 0} \frac{\int_{B_\epsilon(x)} f(t)\, d\mu(t)}{\mu(B_\varepsilon(x))},$$

isso é, o conjunto dos pontos $x \in \mathbb{R}^d$ onde a igualdade não é valida, tem a μ-medida zero.

Seja $x \in \Omega$, e suponha que $\eta(x) \geq 1/2$. Para cada $\epsilon > 0$ bastante pequeno, estritamente mais de metade dos pontos y da bola $B_\epsilon(x)$ têm a propriedade $\eta(y) \geq 1/2$. Em particular, se k é bastante grande e k/n bastante pequeno, então a menor bola ao torno de y que contém exatamente k pontos de uma amostra aleatória de n pontos possui esta propriedade: a maioria dos pontos y da bola tem $\eta(y) \geq 1/2$. Como os k vizinhos mais próximos de x são elementos aleatórios da bola, segundo a lei dos grandes números, a maioria deles possuem a mesma propriedade ($\eta(y) \geq 1/2$) com alta probabilidade, e o voto majoritário associará a x o rótulo 1. Desta maneira, no limite $n \to \infty$, o classificador k-NN associará a μ-quase cada ponto $x \in \Omega$ o mesmo rótulo que o classificador de Stone.

Assim, dentro do modelo atual da aprendizagem estatística, o classificador k-NN, com alta confiança, dará uma resposta correta a longo prazo, quando o tamanho da amostra é bastante grande.

No próximo capítulo analisamos algumas dificuldades relacionadas à dimensão de conjunto de dados.

3 Maldição de dimensionalidade

A dimensão do nosso conjunto de dados para deteção de intrusos na rede é somente 7. Mas existem os conjuntos de dados de uma dimensão muito maior. Por exemplo, a dimensão do conjunto *Phoneme* [20] é 256, é um subconjunto de $\mathbb{R}^{256}$. Um exemplo um pouco extremo é o conjunto de dados do Instituto de Cardiologia da Universidade de Ottawa, com o qual nossa equipe de pesquisa está trabalhando. Os pontos de dados são as sequências genômicas,

$$X \subseteq \{A, T, G, C\}^d,$$

onde a "dimensão" $d \sim 870,000$, enquanto a tamanho do conjunto não é muito grande ($n \sim 4,000$, os dados correspondem aos pacientes individuais).

Para $d \gg 1$, muitos algoritmos conhecidos na ciência de dados muitas vezes levam muito tempo e tornam-se ineficientes. Mesmo em dimensões baixas a médias (tais como 7) os algoritmos tornam-se menos eficientes que em dimensões 1 ou 2. Então, o que está acontecendo nos domínios de alta dimensão?

Consideremos um domínio, Ω, potencialmente de "alta dimensão", como a esfera euclideana:

$$\mathbb{S}^d = \{x \in \mathbb{R}^{d+1} \mid |x| = 1\}.$$

Suponhamos que o único meio de estudar o objecto em questão seja por uma série dos experimentos aleatórios, do seguinte modo. Cada experimento produz um ponto $x \in X$ tirado

de maneira aleatória (cuja a distribução é conforme a medida natural de X, como o volume), e cada vez podemos registrar os valores $f(x)$ de funções (quantidades observáveis)

$$f\colon X \to \mathbb{R}$$

para x. Quanta informação sobre a geometria de X podemos obter desta maneira?

Por exemplo, o que podemos deduzir sobre o *diâmetro* de X? O diâmetro de X é a quantidade

$$\operatorname{diam} X := \sup\{d(x,y) \mid x, y \in X\},$$

onde $d(x,y)$ denota a distância entre x e y. Nesta situação, como as observáveis $f\colon X \to \mathbb{R}$, é lógico considerar as funções *Lipschitz contínuas* com a constante de Lipschitz 1, isso é, as funções que não aumentam a distância:

$$\forall x, y \in X \quad |f(x) - f(y)| \leq d(x,y).$$

Eis uma fonte das tais funções.

Exercício 3.1. *Seja $x_0 \in X$ um ponto de X qualquer. Mostrar que a função distância definida por*

$$x \mapsto \operatorname{dist}(x_0, x)$$

é Lipschitz contínua com a constante 1*.*

Por conseguinte, obtemos o resultado seguinte.

Exercício 3.2. *Mstrar que*

$$\begin{aligned} \operatorname{diam} X \quad = \quad & \sup\{|f(x) - f(y)| : x, y \in X, \\ & f\colon X \to \mathbb{R} \text{ } e \text{ 1-}\textit{Lipschitz contínua}\}. \end{aligned}$$

Se nós pudéssemos medir os valores de *todos as observáveis* para *todos os pares de pontos de* X e então escolher o supremo, saberíamos o diâmetro $\operatorname{diam} X$. Mas isto é impossível. Podemos escolher *uma* observável f, e gerar então a seqüência mais ou menos longa, mas *finita*, de pontos aleatórios,

$$x_1, x_2, \ldots, x_N,$$

registrando cada vez o valor $f(x_i)$, $i = 1, 2, 3, \ldots$.

Depois que produzimos uma série de números reais

$$f(x_1), f(x_2), \ldots, f(x_N),$$

calcularemos a diferença máxima

$$D_N = \max_{i,j=1}^{N} |f(x_i) - f(x_j)| .$$

É imediato que,

$$D_N \leq D,$$

e o que o valor D_{N+1} obtenido na etapa seguinte satisfaz

$$D_N \leq D_{N+1} \leq D,$$

de modo que os valores D_N "melhoram" cada vez.

Pararemos o experimento quando a probabilidade de melhorar o valor precedente, D_N, se torna demasiado pequena. Mais precisamente, seja $\kappa > 0$ (um valor limiar) um número fixo muito pequeno, tal como $\kappa = 10^{-10}$ (sugerido por Gromov).

Nós pararemos depois que o número $D = D_N$ satisfaz

$$\mu\{x \mid |f(x) - M| < D\} > 1 - \kappa\},$$

onde μ é a medida "natural" sobre X. O valor $D = D_N$ obtido da esta maneira chama-se *o diâmetro observável* de X. Mais precisamente, o diâmetro observável obs-diam $_\kappa X$ é definido por

$$\begin{aligned}\text{obs-diam}_{\,\kappa} X \quad = \quad & \inf\{D > 0\text{: para cada observável } f \text{ sobre } X, \\ & \mu\{x \in X \mid |f(x) - M| \geq D\} \leq \kappa\}.\end{aligned}$$

Ilustraremos o conceito para as esferas euclideanas $\mathbb{S}^n$. Neste experimento, substituímos a reta $\mathbb{R}$ com uma "tela" $\mathbb{R}^2$, com a projeção coordenada $\mathbb{R}^{d+1} \to \mathbb{R}^2$,

$$(x_1, x_2, \ldots, x_{d+1}) \to (x_1, x_2),$$

como a observável. O número dos pontos tirandos $N = 1000$. A linha pontilhada representa a projeção da esfera de raio um (o círculo do raio um), enquanto a linha sólida mostra um círculo de tal raio que a probabilidade de um evento de que a projeção de um ponto aleatório na esfera esteja fora deste círculo é menos do que $\kappa = 10^{-10}$. Em outras palavras, o diâmetro do círculo solido é o diâmetro observável da esfera $\mathbb{S}^d$. Veja as Figuras 11, 12 e 13.

É possível provar que o diâmetro observável da esfera satisfaz

$$\text{obs-diam}_{\,\kappa}(\mathbb{S}^d) = O\left(\frac{1}{\sqrt{d}}\right)$$

para cada valor limiar $\kappa > 0$. Em outras palavras, assintoticamente, o diametro observável da esfera $\mathbb{S}^n$ é de ordem $1/\sqrt{d}$.

Come o diâmetro atual da esfera $\mathbb{S}^d$ é 2, uma esfera da alta dimensão aparece como um "cometa" formado de um "nú-

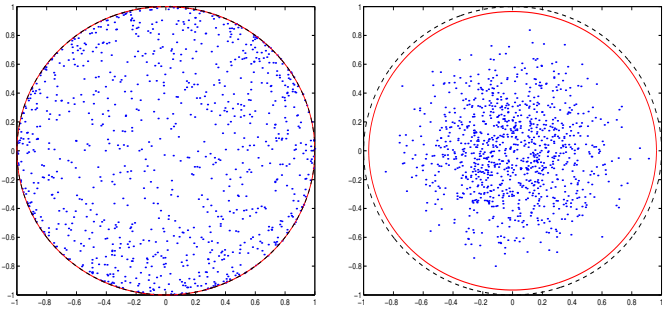

Figura 11 – $\mathbb{S}^2$ e $\mathbb{S}^{10}$

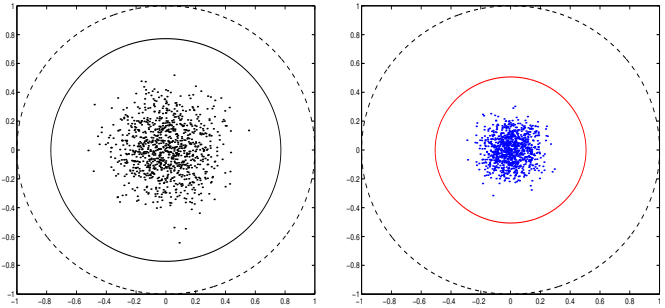

Figura 12 – $\mathbb{S}^{30}$ e $\mathbb{S}^{100}$

cleo" muito pequeno e um "envoltório gasoso" de grande tamanho e de densidade baixa. (Figura 14).

Esta observação é típica de outros objetos geométricos da alta dimensão. Por exemplo, é possível mostrar que o diâmetro observável do cubo unitário,

$$\mathbb{I}^d = \{x \in \mathbb{R}^d \mid \forall i = 1, \ldots, d,\ 0 \leq |x_i| \leq 1\},$$

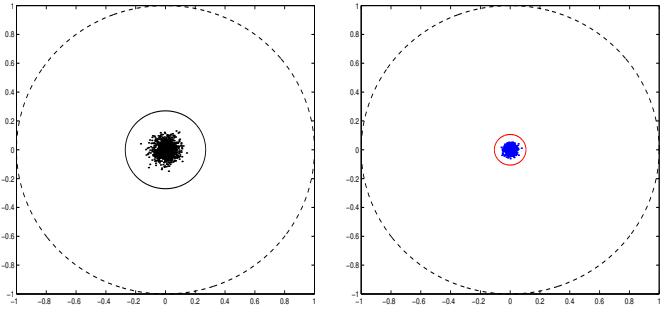

Figura 13 – $\mathbb{S}^{500}$ e $\mathbb{S}^{2500}$

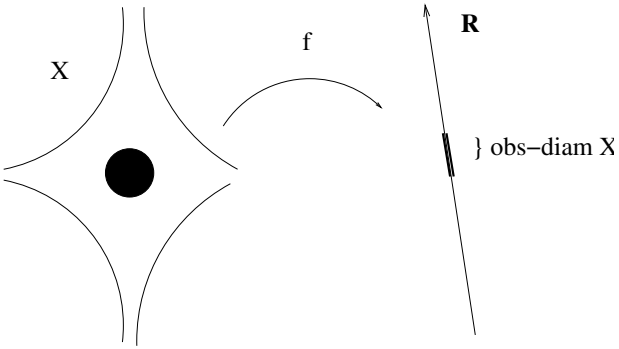

Figura 14 – O diâmetro observável de um espaço da alta dimensão.

satisfaz

$$\text{obs-diam}_{\kappa}(\mathbb{I}^d) = O(1).$$

Isso é, assintoticalemte obs-diam $_{\kappa}(\mathbb{I}^d)$ é constante. Ao mesmo tempo,

$$\text{diam}(\mathbb{I}^d) = \sqrt{d}.$$

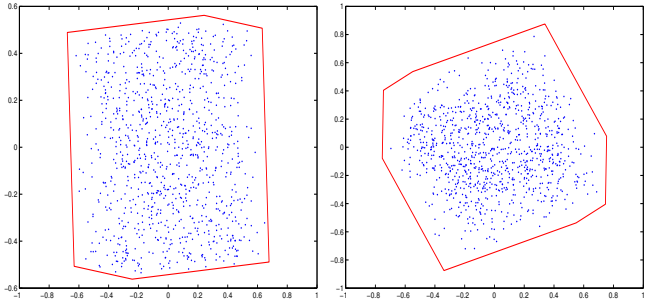

Figura 15 – Projeções do cubo $\mathbb{I}^d$ e dos $1,000$ pontos aleatórios no cubo sobre um plano aleatório, $d = 3, 4$.

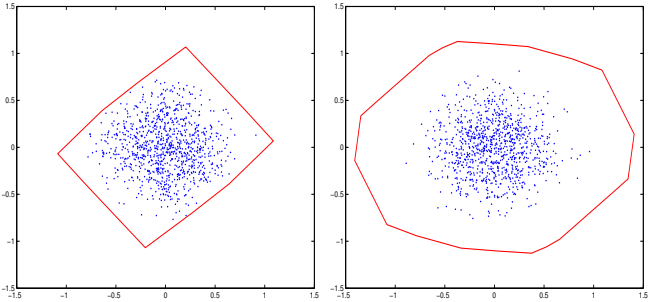

Figura 16 – O mesmo, $d = 5, 10$.

Com efeito, em dimenções altas a projeção ortogonal do cubo $\mathbb{I}^d$ com seus $N = 1000$ pontos aleatórios na direção de um plano aleatório assemelha-se fortemente à projeção da esfera. Veja as Figuras 15, 16, e 18.

A dependência do diâmetro observável no valor limiar κ não é muito sensível (somente logarítmica).

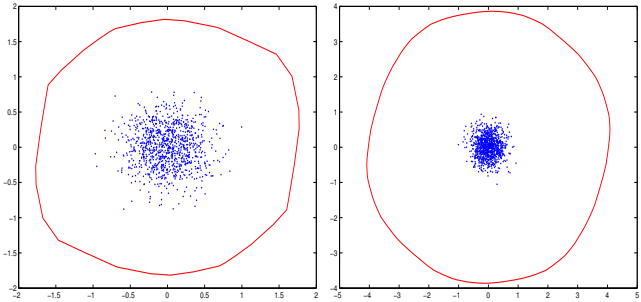

Figura 17 – O mesmo, $d = 20, 100$.

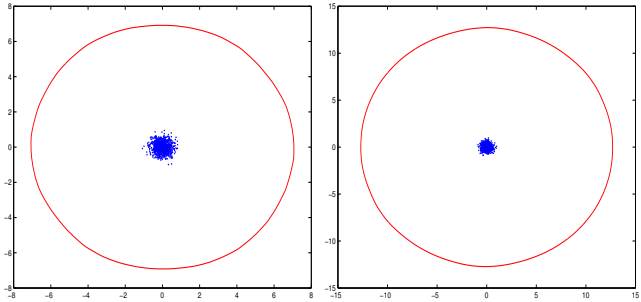

Figura 18 – O mesmo, $d = 300, 1000$.

O *fenômeno de concentração de medida sobre as estruturas de alta dimensão* pode ser exprimido de seguinte maneira informal:

> *O diâmetro observável de um objeto geométrico de alta dimensão é tipicamente demasiado pequeno comparado ao diâmetro atual:*
>
> $$\text{obs-diam}\,(X) \ll \text{diam}\,(X).$$

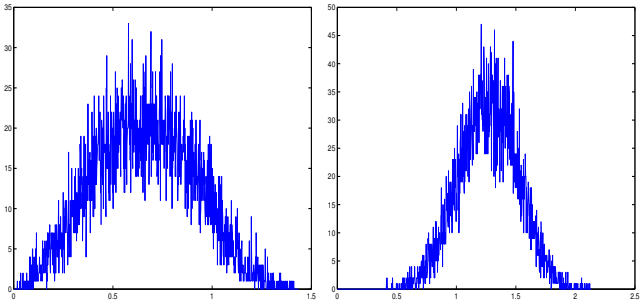

Figura 19 – Distribuição das distâncias entre $10,000$ pontos aleatórios do cubo $\mathbb{I}^d$, $d = 3, 10$.

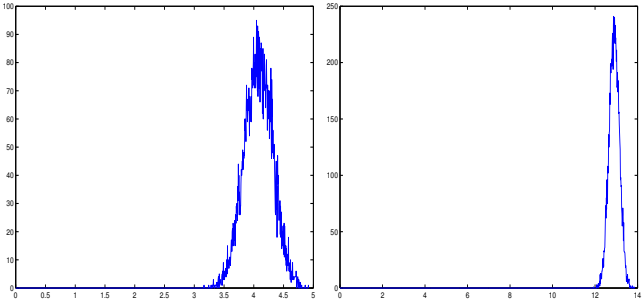

Figura 20 – O mesmo, $d = 100, 1000$.

A formulação mais precisa usa a noção do *tamanho característico* de X em vez do diâmetro. Sobre um espaço de grande dimensão, os valores da distância $d(x, y)$ tipicamente concentram em torno da experança da distância, ou do *tamanho característico* de X,

$$\operatorname{charSize}(X) = \mathbb{E}_{\mu\otimes\mu}(d(x, y)).$$

Veja as Figuras 19 e 20 pelo cubo $\mathbb{I}^d$.

Por exemplo, o tamanho caractéristico da esfera é, assintoticamente, $O(1)$:

$$\text{charSize}\,(\mathbb{S}^n) \to \sqrt{2} \text{ quando } n \to \infty.$$

O fenômeno de concentração de medida na forma mais exata diz o que

Diâmetro observável $\ll$ tamanho característico.

O fenômeno de concentração da medida é o assunto de estudo de uma disciplina matemática relativamente nova: a *análise geométrica assintótica.* Esta introspeção na geometria dos objectos de dimensão alta é a mais importante, e tem muitas aplicações e conseqüências amplas em ciências matemáticas.

O que o fenômeno significa no contexto concreto de classificador k-NN? Eis uma reformulação heurística equivalente (embora não evidente) do fenômeno:

> *Tipicamente, num espaço Ω de grande dimensão, para cada subconjunto $A \subseteq \Omega$ que contém pelo menos a metade dos pontos, a maior parte dos pontos de Ω estão próximos ao A.*

Formalizamos a noção de uma "estrutura". Seja $\Omega = (\Omega, \rho, \mu)$ um espaço métrico, munido de uma medida de probabilidade μ. Consideremos 4 exemplos.

(1) Seja $d \in \mathbb{N}$. O *cubo de Hamming* de dimensão d é a coleção de seqüências de d digitos $0-1$ (palavras binárias de comprimento d). Designamos-o $\{0,1\}^d$ ou Σ^d. Assim, um elemento

típico $\sigma \in \Sigma^d$ é da forma

$$\sigma = \sigma_1 \sigma_2 \cdots \sigma_d,$$

ou $\sigma_i \in \{0, 1\}$ para todo i. A *distância de Hamming normalizada* entre duas n-palavras $\sigma, \tau \in \Sigma^d$ é definida para

$$d(\sigma, \tau) = \frac{1}{d} \sharp\{i \colon \sigma_i \neq \tau_i\}.$$

Seja A um subconjuinto qualquer de Σ^d. O valor da *medida uniforme normalizada* de A é dada por

$$\mu_\sharp(A) = \frac{|A|}{2^d}.$$

(2) A esfera euclideana unitária $\mathbb{S}^d$ admite duas métricas padrão: a distância euclideana induzida de $\ell^2(d+1)$,

$$d_{eucl}(x, y) = \| x - y \|_2,$$

e a distância geodésica, em outros termos, o ângulo entre dois vetores:

$$d_{geo}(x, y) = \angle(x, y).$$

As duas distâncias são equivalentes: qualquer que sejam $x, y \in \mathbb{S}^d$, temos

$$d_{eucl}(x, y) \leq d_{geo}(x, y) \leq \frac{\pi}{2} d_{eucl}(x, y),$$

e no caso onde $d_{geo}(x, y) \leq \pi/2$, temos

$$d_{eucl}(x, y) \leq d_{geo}(x, y) \leq \frac{\pi}{2\sqrt{2}} d_{eucl}(x, y). \tag{3.1}$$

O grupo ortogonal

$$O(d) = \{u \in M_d(\mathbb{R}) \colon u^t u = u u^t = 1\}$$

age sobre a esfera pelas isometrias

$$\mathbb{S}^d \ni x \mapsto ux \in \mathbb{S}^d, \quad u \in O(d).$$

Existe uma única medida de probabilidade boreliana $\nu = \nu_d$ sobre $\mathbb{S}^d$ invariante sobre isometrias, isto é, tal que

$$\nu_d(A) = \nu_d(uA) \text{ para toda } u \in O(d).$$

A medida ν chama-se a *medida de Haar*. Se λ_d denota a medida de Lebesgue no espaço $\mathbb{R}^d$, então para cada sub-conjunto boreleano $A \subseteq \mathbb{S}^d$ temos

$$\nu_d(A) = \frac{\lambda_{d+1}(\tilde{A})}{\lambda_{d+1}(B_{d+1})},$$

onde $\tilde{A}$ é o cone sobre A:

$$\tilde{A} = \{ta \colon t \in [0,1],\ a \in A\}$$

e B_d é a bola fechada do raio um no espaço euclideano $\ell^2(d)$.

(3) Os espaço euclideano $\mathbb{R}^d$ munido da medida gaussiana γ_d.

(4) O cubo $[0,1]^d$ munido da medida uniforme.

Denotaremos

$$A_\epsilon = \{x \in \Omega \colon \exists a \in A \ \ \rho(x,a) < \epsilon\}$$

a ϵ-vizinhança do sub-conjunto A de Ω.

Definicão 3.3. *Seja* $(\Omega_d, \rho_d, \mu_d)$, $d = 1, 2, 3, \ldots$ *uma família de espaços métricos munidos de uma medida de probabilidade boreleana (espaços métricos com medida). Esta família é* uma

família de Lévy *se, para cada família* A_d, $d = 1, 2, \ldots$, *de subconjuntos boreleanos de* Ω_d, *tais que*

$$\liminf \mu_d(A_d) > 0,$$

e por cado $\epsilon > 0$, *temos*

$$\mu_d((A_d)_\epsilon) \to 1.$$

As famílias "naturais" dos espaços métricos com medida são tipicamente as famílias de Lévy. Tais são os exemplos em (1) (o resultado é conhecido na teoria de informação como o "Blowing-Up Lemma") e (2) (Paul Lévy, 1922). Os espaços em (3) e (4) não formam as famílias de Lévy, mas eles transformam-se em famílias de Lévy após uma *renormalização* pelo fator inverso ao tamanho caraterístico. O tamanho caraterítico de $(\mathbb{R}^d, \gamma_d)$ e de $[0,1]^d$ com a medida uniforme é do ordem $O(\sqrt{d})$, e se a distância nestes espaços é multiplicada pelo fator $1/\sqrt{d}$, as famílias resultantes são as de Lévy.

Um instrumento conveniente para quantificar o fenómeno da concentração de medida é a *função de concentração.*

Definicão 3.4. *Seja* (Ω, d, μ) *um espaço métrico com medida. A* função de concentração *de* Ω, *notada* $\alpha_\Omega(\epsilon)$, *é definida pelas condições seguintes:*

$$\alpha(\epsilon) = \begin{cases} \frac{1}{2}, & \text{se } \epsilon = 0, \\ 1 - \min\left\{\mu_\sharp\left(A_\epsilon\right) : A \subseteq \Sigma^n, \quad \mu_\sharp(A) \geq \frac{1}{2}\right\}, & \text{se } \epsilon > 0. \end{cases}$$

Exercício 3.5. *Mostrar que*

$$\alpha(\Omega, \epsilon) \to 0 \text{ quando } \epsilon \to \infty.$$

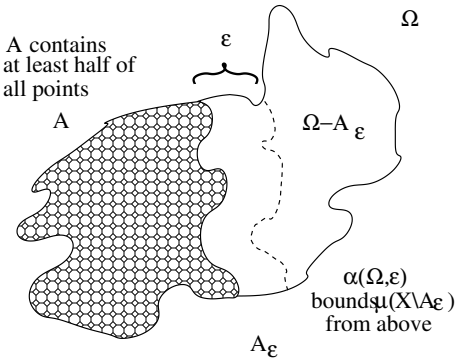

Figura 21 – A função de concentração $\alpha(\Omega, \epsilon)$.

Teorema 3.6. *Uma família* $(\Omega_d, \rho_d, \mu_d)$ *é uma família de Lévy se e apenas se as funções de concentração tendem a zero:*

$$\alpha(\Omega_d, \epsilon) \to 0 \text{ para cada } \epsilon > 0.$$

□

Definicão 3.7. *Uma família de Lévy* (Ω_d, d_d, μ_d) *é chamada uma* família de Lévy normal *se existem* $C_1, C_2 > 0$ *tais que*

$$\alpha(\Omega_d, \epsilon) \leq C_1 e^{-C_2 \epsilon^2 d}.$$

Teorema 3.8. *Por a função de concentração do cubo de Hamming* Σ^d *temos*

$$\alpha_{\Sigma^d}(\epsilon) \leq 2e^{-\epsilon^2 d/2}.$$

□

Aqui está a ligação com o diâmetro observável: sobre uma estrutura de grande dimensão, toda função Lipschitz contínua é quasi constante em toda parte exceto sobre um conjunto da medida muito pequena.

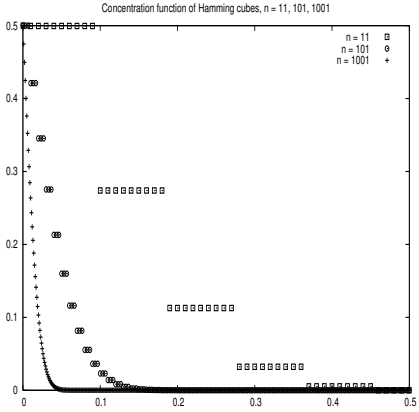

Figura 22 – As funçãos de concentração dos cubos de Hamming por $d = 11, 101, 1001$.

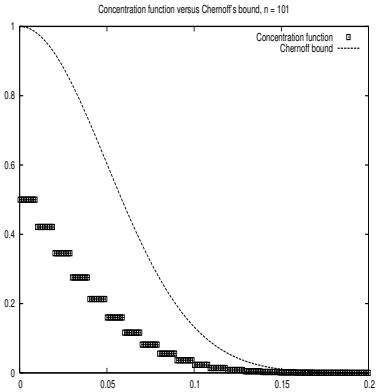

Figura 23 – Função de concentração do cubo de Hamming Σ^{101} e a cota superior gaussiana de Chernoff para os valores pequenos de ϵ

Relembramos que um número real $M = M_f$ é dito um *valor mediano* de uma função boreliana f, sobre um espaço com medida de probabilidade (Ω, μ) se

$$\mu\{x \in \Omega : f(x) \geq M\} \geq \frac{1}{2} \text{ e } \mu\{x \in \Omega : f(x) \leq M\} \geq \frac{1}{2}.$$

Um valor mediano $M = M_f$ existe sempre, mas geralmente, não é único.

Exercício 3.9. *Seja f uma função Lipschitz contínua com a constante de Lipschitz $L \geq 0$ sobre um espaço métrico com medida, (Ω, ρ, μ). Provar que*

$$\mu\{|f(x) - M| > \epsilon\} \leq 2\alpha_\Omega\left(\frac{\epsilon}{L}\right).$$

Mais geralmente, se f é uniformemente contínua de tal modo que

$$\forall x, y \in X, \quad d(x, y) < \delta \Rightarrow |fx - fy| < \epsilon,$$

então

$$\mu\{|f(x) - M| > \epsilon\} \leq 2\alpha_X(\delta).$$

▲

Exercício 3.10. *Deduzir a lei dos grandes números do teorema 3.8, aplicando o exercício 3.9 à função real*

$$f(\sigma) = \frac{1}{d}\sum_{i=1}^{d} \sigma_i$$

sobre o cubo de Hamming. (É por isso que o teorema 3.8 é as vezes chamado o lei geométrica dos grandes números.*)*

A função da distância $d(-, p)$ de um ponto p fixo qualquer é Lipschitz contínua (com a constante 1), e em domínios

de dimensão alta uma tal função concentra-se em torno do valor mediano. Este efeito pronuncia-se já em dimensões médias, tais como $d = 14$ na Figura 24.

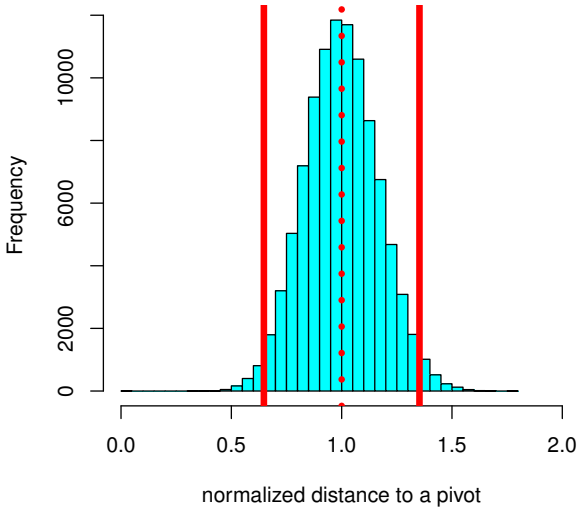

Figura 24 – Histograma das distâncias a um ponto escolhido aleatoriamente em um conjnto de dados X com $n = 10^5$ points, tirados de uma distribuição gaussiana em $\mathbb{R}^{14}$.

Em consequência, a distância média $\mathbb{E}(\epsilon_{NN})$ de um ponto do domínio ao seu vizinho mais próximo na amostra aleatória é quase igual ao tamanho cataterístico do domínio (isso é, a distância média entre dois pontos do domínio), quando a dimensão d vai para o infinito, desde que o tamanho da amostra, n, cresce de maneira subexponencial em d (o que é sempre o caso). Veja

Figura 25.

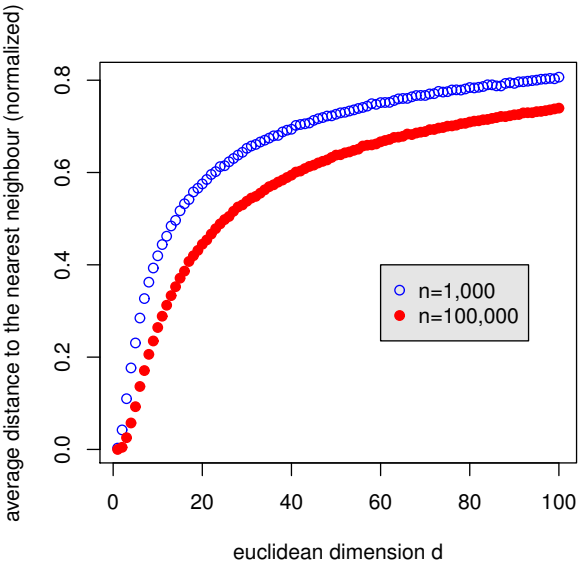

Figura 25 – A relação entre a distância média para o vizinho mais próximo e o tamanho caraterístico em um conjunto de n pontos tirados aleatoriamente de uma distribuição gaussiana em $\mathbb{R}^d$.

Também, na Figura 24, as linhas verticais marcam a distância média normalizada $1 \pm \epsilon_{NN}$, onde ϵ_{NN} é a distância média do vizinho mais próximo.

Isso é conhecido na ciência de dados como o *paradoxo de espaço vazio*. Este paradóxo tem uma consequência imediata para o classificador k-NN. Seja $x \in \Omega$ um ponto qualquer. Denotameros $\varepsilon_{NN}(x)$ a distância de x para o seu vizinho mais

próximo na amostra aleatória, X. Na consequência do paradoxo de espaço vazio, uma grande quantidade de pontos de X estão quase a mesma distância de x que o seu vizinho mais próximo. Mais formalmente, seja $c > 0$, e dizemos, seguindo [2], que a consulta de vizinho mais próximo de x é *c-instável* se a bola do raio $(1 + c)\varepsilon_{NN}(x)$ centrada em x contém pelo menos metade dos pontos de X. (Figura 26.)

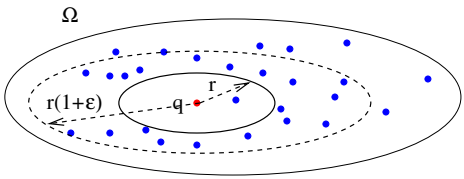

Figura 26 – Instavilidade da busca do vizinho mais próximo.

Usando a concentração da medida, não é difícil de monstrar que, pelo $c > 0$ fixo, no limite $d \to \infty$ a maioria das buscas serão c-instáveis.

Nas dimensões baixas, o fenômeno está fraco (Figura 27, a esquerda, o conjunto de dados *Segment* da *UCI data repository* [27]), mas nas dimensões médias, é já pronunciado (Figura 27, a direita, o subconjunto aleatório da distribuição gaussiana em $\mathbb{R}^{14}$). Aqui, $k = 20$ e $c = 0,5$. A linha esquerda vertical corresponde ao valor médio do raio da bola que contém k vizinhos mais próximos, $\varepsilon_{k\text{-NN}}$, e a segunda linha corresponde a $(1 + c)\varepsilon_{k\text{-NN}}$. Para o conjunto *Segment,* a segunda bola contém em média 60 pontos. Para o gaussiano, o valor correspondente já é de $1,742$ pontos.

O fenômeno da instabilidade significa uma perda ób-

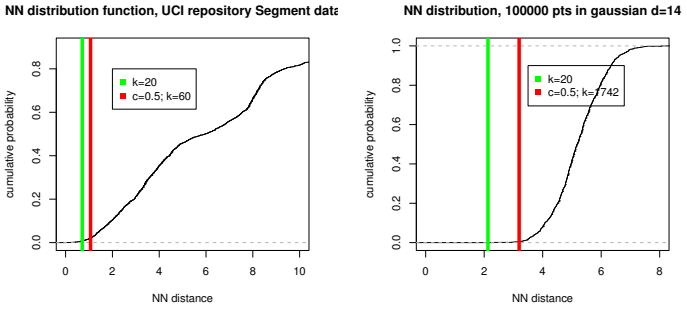

Figura 27 – A fração média dos pontos de dados nas bolas de raio $(1+c)\varepsilon_{k\text{-NN}}$.

via da importância do fato de ser o vizinho mais próximo. Por exemplo, no caso de um erro quase inevitável de recuperação do vizinho mais próximo exato, o rótulo do vizinho substituido será mais ou menos aleatório. O desempenho do classificador k-NN (e de qualquer outro algoritmo baseado nos vizinhos mais próximos) degrada especialmente em dimensões altas, mas também em dimensões médias, mesmo se não tão notóriamente. No próximo capítulo, vamos discutir algumas receitas contra essa *maldição da dimensionalidade.*

Entre os livros tratando o fenômeno de concentração de medida, [17] é o mais acessível, [13] o mais abrangente e [10] contém uma riqueza de idéias.

4 Redução de dimensionalidade

Seja Ω um domínio, contendo uma amostra σ. *Redução de dimensionalidade* significa escolhendo uma função $f: \Omega \to W$ de Ω para um domínio W de dimensão mais baixa. A fim de classificar um ponto de dados $x \in \Omega$, vamos aplicar um algoritmo de classificação no espaço W ao ponto $f(x)$ e a amostra $f(\sigma)$, esperando que o desempenho do algoritmo em W seja mais eficaz, e que a função f conserve a estrutura geral e os padrões presentes no conjunto de dados X. Existem muitas métodos da redução de dimensionalidade.

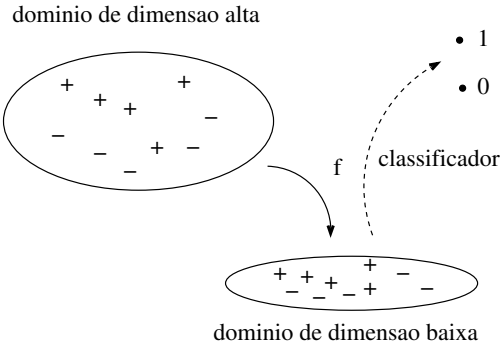

Figura 28 – Redução de dimenionalidade

4.1 PCA (Principal Component Analysis)

Este algoritmo padrão é o mais antigo e provavelmente o mais comum na ciência de dados. Aqui, entre as coordenadas

(caraterísticas) do espaço $\Omega = \mathbb{R}^d$, escolhemos as coordenadas mais importantes do ponto de visto da geometria do conjunto de dados. O algoritmo PCA foi implementado em R.

Exercício 4.1. *Estudar o algoritmo PCA [23, 3] e aplicar ao conjunto de dados* Phoneme*, seguido pelo classificador k-NN num espaço de uma dimensão menor. Conseguiu melhorar a precisão?*

4.2 Projeções aleatórias (Lema de Johnson–Lindenstrauss)

Este método relativamente recente é uma aplicação da concentração de medida.

Theorem 4.2 (Lema de Johnson–Lindenstrauss [11])**.** *Seja X um sub-conjunto com n elementos num espaço de Hilbert $\mathcal{H}$, e seja $0 < \epsilon \leq 1$. Então existe um operador linear $T\colon \mathcal{H} \to \ell^2(k)$, onde*

$$k = O(\epsilon^{-2} \log n),$$

tal que

$$(1-\epsilon)\| x-y \| < \| T(x) - T(y) \| < (1+\epsilon)\| x-y \|$$

para todos $x, y \in X$.

Obviamente, sem perda de generalidade, podemos supor que $\dim \mathcal{H} = n$. A dimensão do espaço reduzido, $\ell^2(k)$, é *logarítmica* em n. Por exemplo, se $n \geq 158$, então pode-se mostrar que a dimensão k satisfaz

$$k \leq \left\lceil \frac{17 \log n}{\epsilon^2} \right\rceil .$$

(Claro, os limites podem ser melhoradas).

Temos uma consequância particularmente interessante do resultado. Se um espaço de Hilbert, $\mathcal{H}$, contém um sistema ortonormal de n vetores, então, claro, sua dimensão deve ser pelo menos n. No entanto, chamamos um sistema de vetores de norma 1 ϵ-*quase ortonormal* se o ângulo entre quaisquer dois vetores distintos é $\pi/2 \pm \epsilon$. O resultado acima implica a existência de sistemas quase ortonormais de vetores cujo tamanho é *exponencial* na dimensão do espaço de Hilbert.

Como escolher o operador T? Se $\dim \mathcal{H} = n$, então T é dado por uma matriz do tamanho $k \times n$. Uma coisa interessante que torna o lema de Johnson–Lindenstrauss altamente aplicável na prática de computação, é que os coefficientes da matriz de T podem ser escolhidos aleatoriamente, como uma sequência dos reais independentes identicamente distribuidos, seguindo, por exemplo, a distribuição gaussiana, ou mesmo a distribuição de Bernoulli. As duas boas referências são os livros [15] e [29].

Exercício 4.3. *Impementar o algoritmo dos projeções aleatórias em R e aplicar ao conjunto de dados* Phoneme*, combinando com o classificador k-NN.*

4.3 Redução de dimensionalidade usando injeções borelianas

Conceitos básicos de teoria descritiva dos conjuntos [12] oferecem uma nova abordagem para a redução de dimensionalidade no contexto da aprendizagem automática estatística, sugerido em [19] e aplicada com sucesso na competição CDMC'2013 pela equipe consistente do ministrante e de três alunos: Gaël Giordano, Hubert Duan, e Stan Hatko.

Geralmente assumimos que as aplicações f que realizam a redução de dimensionalidade são contínuas, até mesmo Lipschitz contínuas. Esta é uma condição muito restritiva. No entanto, examinando o modelo teórico existente que estabelece uma base para a aprendizagem estatística, pode notar-se que o teorema de Stone é na verdade insensía *estrutura euclidiana* (ou seja, estrutura métrica ou mesmo topológica) no domínio, enquanto a estrutura *boreliana* permanece intacta. Isto permite, através de um isomorfismo boreliano (o mesmo uma injeção boreliana), reduzir os dados para um caso de baixa dimensão, até mesmo unidimensional, após o qual o algoritmo k-NN continua a ser universalmente consistente.

Definicão 4.4. *A* sigma-álgebra *de subconjuntos de um conjunto* Ω *é uma família (não vazia)* $\mathscr{A} \subseteq 2^{\Omega}$ *com as propriedades:*

1. *Se* $A_1, A_2, \ldots, A_n, \ldots$ *pertençam a* $\mathscr{A}$*, então* $\cup_{i=1}^{\infty} A_i \in \mathscr{A}$*.*
2. *Se* $A \in \mathscr{A}$*, então* $\Omega \setminus A \in \mathscr{A}$*.*

Um conjunto Ω munido de uma sigma-álgebra se chama um *espaço mensurável.* Se Ω é um espaço métrico, denotaremos $\mathcal{B}_X$ a menor sigma-álgebra que contém todos as abertas de Ω. Esta sigma-álgebra $\mathcal{B}_X$ é a *estrutura boreliana* de Ω, e os elementos de $\mathcal{B}_X$ são os *conjuntos borelianos.* Se o espaço métrico Ω é separável e completo, sua estrutura boreliana é dita *estrutura boreliana padrão,* e o espaço mensurável $(\Omega, \mathcal{B}_X)$ é dito um *espaco boreliano padrão.* Neste caso, a estruture boreliana é gerada por todas as bolas abertas $B_r(x)$, $x \in \Omega$, $r > 0$.

Uma aplicação $f{:}\,\Omega \to W$ entre dois espacos borelianos é dita *isomorfismo boreliano* se f é bijetiva, e f e f^{-1} são

borelianas. Isso significa que f estabelece uma bijeição entre a estrutura boreliana $\mathcal{B}_\Omega$ e $\mathcal{B}_W$.

Cada aplicação contínua é boreliano, e cada homeomorfismo (isso é, uma bijeição contínua, com o inverso contínuo) é um isomorfismo boreliano. Mas existem muito mais aplicações borelianas que aplicações contínuas, e muito mais isomorfismos bolerianos que homeomorfismos.

Por exemplo, é bem conhecido e facilmente mostrado que como espaços topológicos, o intervalo $[0,1]$ e o quadrado $[0,1]^2$ não são homeomorfos. Ainda mais, não há nenhum injeção contínua de $[0,1]^2$ para $[0,1]$. Ao mesmo tempo, existe uma injeção boreliana do quadrado no intervalo. Ela pode ser obtido usando o "entrelaçamento" dos digitos nas expansões binárias de x e de y num par $(x,y) \in [0,1]^2$ (sujeito as precauções habituais sobre as seqüências infinitas de uns):

$$[0,1]^2 \ni (0.a_1a_2\ldots, 0.b_1b_2\ldots) \mapsto (0.a_1b_1a_2b_2\ldots) \in [0,1]. \quad (4.1)$$

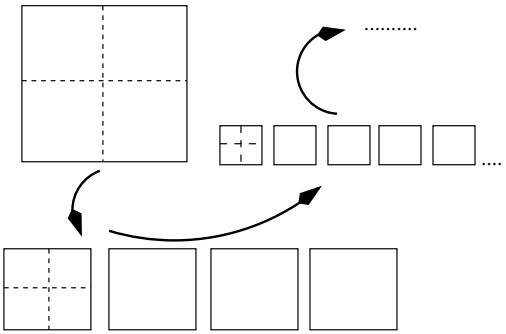

Figura 29 – Construindo um isomorfismo boreliano entre o quadrado e um intervalo.

Por uma representação geométrica desta injeção, veja a figura 29. A aplicação f acima não é surjetiva, por exemplo o ponto $0.10101010\ldots$ não é na imagem de f. Mas ela pode ser modificada a fim de obter um isomorfismo boreliano entre $[0,1]^2$ e $[0,1]$. No lugar da base 2, pode ser uma base qualquer.

Esta construção pode ser generalizada para mostrar que não há muita diversidade entre os espaços borelianos padrão. Eis um resultado clássico.

Teorema 4.5. *Sejam Ω e W dois espaços métricos separáveis e completos, da cardinalidade $\mathfrak{c} = 2^{\aleph_0}$ cadaum. (Por exemplo, isso é o caso se eles não contém os pontos isolados). Então os espaços borelianos correspondentes são isomorfos.* □

Este será o caso da maioria dos domínios de interesse na teoria. Por exemplo, o conjunto de Cantor, o intervalo unitário, o espaço euclidiano $\mathbb{R}^d$, o espaço de Hilbert separável de dimensão infinita ℓ^2, e na verdade todos espaços de Fréchet separáveis não triviais são todos isomorfos entre eles como espaços borelianos. Sua estrutura de Borel é a mesma do *espaco de Borel padrão* com cardinalidade de contínuo.

Agora, seja $(\Omega, \mathscr{B})$ um espaço boreliano padrão (um domínio), e seja μ uma medida de probabilidade sobre $\Omega \times \{0,1\}$, isso é, uma aplicação

$$\mu\colon \mathscr{B}_{\Omega\times\{0,1\}} \to [0,1],$$

satisfando as propriedades (P1) e (P2) acima.

Se π é a projeição de $\Omega \times \{0,1\}$ pela primeira coordenada,

$$\pi(x,\varepsilon) = x,$$

então a imagem direita da medida μ é uma medida de probabilidade, ν, sobre Ω: se A é um conjunto boreliano em Ω,

$$\nu(A) = \mu(\pi^{-1}(A)).$$

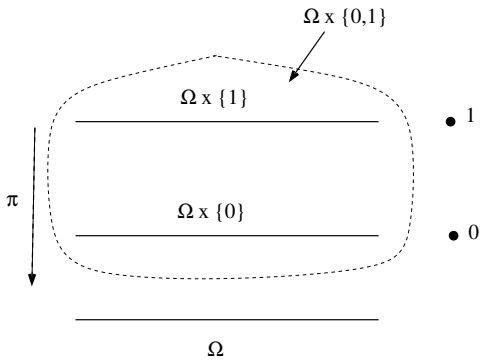

Figura 30 – $\pi\colon \Omega \times \{0,1\} \to \Omega$

A função de regressão, η, é definida pelas condições: se $A \subseteq \Omega$, então

$$\mu(A \times \{1\}) = \int_A \eta(x)\, d\mu(x),$$

$$\mu(A \times \{0\}) = \int_A (1 - \eta(x))\, d\mu(x).$$

Relembramos que o classificador de Bayes (um classificador cujo erro de classificação é o mínimo possivel) é dado por

$$T_{bayes}(x) = \begin{cases} 0, & \text{se } \eta(x) < \frac{1}{2}, \\ 1, & \text{se } \eta(x) \geq \frac{1}{2}. \end{cases}$$

(Veja Figura 31).

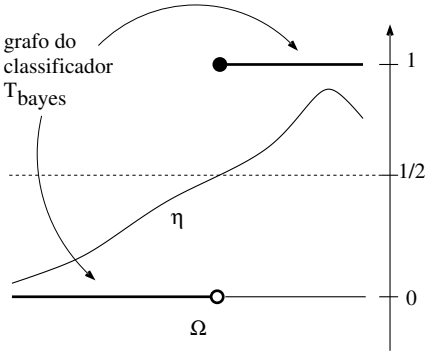

Figura 31 – Função de regressão η e o classificador de Bayes T_{bayes}.

Seja W um outro espaço métrico, e seja $f\colon \Omega \to W$ uma injeção boreliana. Esta f pode ser prolongada até uma injeção boreliana de $\Omega \times \{0,1\}$ em $W \times \{0,1\}$ pela formula óbvia:

$$f(x,\varepsilon) = (f(x),\varepsilon),$$

onde $\varepsilon \in \{0,1\}$. Vamos usar a mesma letra f pela prolongação. Definiremos a imagem direita $f_*\mu$ da medida μ ao longo de f: qual quer seja um boreliano $B \subseteq W \times \{0,1\}$,

$$(f_*\mu)(B) = \mu(f^{-1}(B)).$$

É uma medida de probabilidade borealiana sobre $W \times \{0,1\}$.

Pode-se mostrar sem dificuldade que si

$$X_1, X_2, \ldots, X_n, \ldots$$

é uma sequência das variáveis aleatórias independentes com valores em $\Omega \times \{0,1\}$ segundo a lei μ, então

$$f(X_1), f(X_2), \ldots, f(X_n), \ldots$$

é uma sequência das variáveis aleatórias independentes com valores em $W \times \{0,1\}$ seguindo a lei $f_*(\mu)$.

A medida $f_*(\mu)$ sobre $W \times \{0,1\}$ possui sua própria função de regressão, θ. Não é difícil de verificar que, com efeito,

$$\eta = \theta \circ f.$$

Por consequinte, o classificador de Bayes ("o melhor classificador imaginável") para Ω, T^{Ω}_{bayes}, e o classificador de Bayes para W, T^{W}_{bayes}, satisfazem:

$$\forall x \in \Omega, \quad T^{\Omega}_{bayes}(x) = T^{W}_{bayes}(f(x)).$$

Suponha agora que $\mathcal{L}$ é um classificador universalmente consistente qualquer no domínio W. Definiremos um novo classificador, $\mathcal{L}^f$, como a composisão de $\mathcal{L}$ com a injeção boreliana f:

$$\mathcal{L}^f_n(\sigma)(x) = \mathcal{L}_n(f(\sigma))(f(x)).$$

(Como na Figura 28).

Quando $n \to \infty$, os predições do classificador $\mathcal{L}(f(\sigma))$ no ponto $f(x)$, $x \in \Omega$ aproximam-se das predições do classificador de Bayes em W no ponto $f(x)$. Por conseguinte, as predições do classificador "composto", $\mathcal{L}^f(\sigma)$, aproximam-se das predições do classificador de Bayes em Ω no ponto x. Isso significa que $\mathcal{L}^f$ é universalmente consistente no domínio Ω. O isomorfismo f é uma redução de dimensionalidade que conserva a consistência universal dos algoritmos de aprendizagem supervisionada.

Obtemos o seguinte resultado, que oferece uma nova perspectiva da redução de dimensionalidade em teoria da aprendizagem automática estatística.

Theorem 4.6. *Sejam Ω e W dois domínios (espaços borelianos padrão), e seja $f\colon \Omega \to W$ uma injeção boreliana. Seja $\mathcal{L}$ é um classificador universalmente consistente em W. Então o classificador $\mathcal{L}^f$, obtido pela redução de dimensionalidade f de Ω para W, seguida da aplicação do classificador $\mathcal{L}$, é universalmente consistente em Ω também.*

Em particular, há sempre uma redução de Borel isomórfica do problema em $\mathbb{R}^d$ (ou mesmo num espaço de dimensão infinita) para o caso $d = 1$. As experiências até agora mostram que os melhores resultados são obtidos quando a dimensão é reduzida por um fator constante (por exemplo, entre 4 e 7), que depende do conjunto de dados.

A redução de dimensionalidade Borel isomórfica foi usada com sucesso na competição CDMC'2013, onde o erro de classificação pelo problema de deteção de intrusos numa rede foi reduzido até 0.1 por cento.

Exercício 4.7. *Escrever o código em R para redução de dimensionalidade usando as injeções borelianas, e combinar-o com o classificador k-NN para melhorar o erro de classificação no problema do reconhecimento de voz (o conjunto de dados* Phoneme*). Tentar as bases diferentes de expansão dos números.*

Leitura sugerida

Boas fontes teóricas para alunos de matemática dispostos para aprender o assunto são [1, 8, 16, 28, 30], combinados com a programação prática, por exemplo, após as linhas de [26].

Agradecimentos

Sou grato aos membros da equipe CDMC'2013, particularmente Stan Hatko avec quem trabalhavam sobre a deteção de intrusos atraves da redução Borel isomórfica, e a Professora Maria Inez Cardoso Gonçalvez, por sua ajuda com meu Português ruim.

Referências

[1] Martin Anthony and Peter Bartlett, Neural network learning: theoretical foundations, Cambridge University Press, Cambridge, 1999. xiv+389 pp. ISBN: 0-521-57353-X

[2] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, *When is "nearest neighbor" meaningful?,* in: Proc. 7-th Intern. Conf. on Database Theory (ICDT-99), Jerusalem, pp. 217–235, 1999.

[3] Ed Boone, *PCA in R,* http://www.youtube.com/watch?v=Heh7Nv4qimU

[4] F. Cérou and A. Guyader, *Nearest neighbor classification in infinite dimension,* ESAIM Probab. Stat. **10** (2006), 340–355.

[5] M.J. Crawley, *The R Book,* http://users.humboldt.edu/ygkim/CrawleyMJ_TheRBook.pdf

[6] DARPA Intrusion Detection Data Sets, MIT Lincoln Lab, http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/data/

[7] L. Devroye, *On the almost everywhere convergence of nonparametric regression function estimates,* Ann. Statist. **9** (1981), 1310–1319.

[8] Luc Devroye, László Györfi and Gábor Lugosi, A Probabilistic Theory of Pattern Recognition, Springer-Verlag, New York, 1996. ISBN 0-387-94618-7.

[9] FNN (Fast Nearest Neighbor Search Algorithms and Applications) package, *http://cran.r-project.org/web/packages/FNN/FNN.pdf*

[10] M. Gromov, *Metric Structures for Riemannian and Non-Riemannian Spaces,* Progress in Mathematics **152**, Birkhauser Verlag, 1999.

[11] W.B. Johnson and J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space, Contemp. Math.* **26** (1984), 189–206.

[12] A.S. Kechris, *Classical Descriptive Set Theory,* Springer-Verlag, 1995.

[13] M. Ledoux, *The concentration of measure phenomenon.* Math. Surveys and Monographs, **89**, Amer. Math. Soc., 2001.

[14] J.H. Maindonald, *Using R for Data Analysis and Graphics. Introduction, Code and Commentary,* http://cran.r-project.org/doc/contrib/usingR.pdf

[15] J. Matoušek, *On variants of the Johnson-Lindenstrauss lemma,* Random Structures Algorithms **33** (2008), 142–156.

[16] Shahar Mendelson, *A few notes on statistical learning theory,* In: Advanced Lectures in Machine Learning, (S. Mendelson, A.J. Smola Eds), LNCS 2600, pp. 1-40, Springer 2003.

[17] V.D. Milman and G. Schechtman, *Asymptotic theory of finite-dimensional normed spaces (with an Appendix by M. Gromov)*, Lecture Notes in Math., **1200**, Springer, 1986.

[18] W.J. Owen, *The R guide,* http://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf

[19] V. Pestov, *Is the k-NN classifier in high dimensions affected by the curse of dimensionality?* Computers & Mathematics with Applications **65** (2013), 1427–1437.

[20] Phoneme dataset, http://statweb.stanford.edu/∼tibs/ElemStatLearn/datasets/phoneme.data

[21] D. Preiss, *Gaussian measures and the density theorem,* Comment. Math. Univ. Carolin. **22** (1981), 181–193.

[22] The R project for statistical computing, http://www.r-project.org/

[23] L.I. Smith, A tutorial on Principal Component Analysis, http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

[24] J. Song, H. Takakura and Y. Kwon, *A Generalized Feature Extraction Scheme to Detect 0-Day Attacks via IDS Alerts,* in: The 2008 Inter. Symposium on Applications and the Internet (SAINT2008), IEEE CS Press, 51–56, Turku, FINLAND, 28 July - 1 Aug. 2008.

[25] C. Stone, *Consistent nonparametric regression,* Annals of Statistics **5** (1977), 595–645.

[26] Luis Torgo, *Data Mining with R: Learning with Case Studies,* Chapman & Hall/SRC, 2010.

[27] UCI Machine Learning Data Set Repository, http://archive.ics.uci.edu/ml/

[28] Vladimir N. Vapnik, Statistical learning theory, John Wiley & Sons, Inc., New York, 1998.

[29] S.S. Vempala, *The random projection method,* DIMACS Series in Discrete Mathematics and Theoretical Computer Science, **65**. American Mathematical Society, Providence, RI, 2004.

[30] M. Vidyasagar, Learning and Generalization, with Applications to Neural Networks, 2nd Ed., Springer-Verlag, 2003.