3
Redução da dimensionalidade

Em muitos problemas, o conjunto de dados apresenta muitos atributos, isto é, possuem uma dimensionalidade alta. No entanto, há motivos para se acreditar que os dados pertençam a uma variedade de baixa dimensão, isto é, características ou parâmetros podem ser dependentes entre si, ou em outras palavras, a dimensionalidade dos dados pode ser maior do que a necessária, como no caso de variáveis altamente correlacionadas. Assim, o grande conjunto de parâmetros poderia ser reduzido a um conjunto menor, seja pela eliminação de variáveis consideradas irrelevantes ou pela transformação das variáveis, aliada a preservação de características importantes do conjunto de dados inicial. Em todos os casos, reduzir a dimensionalidade dos dados resulta em uma representação mais eficiente dos mesmos.

As transformações dos dados observados podem seguir dois objetivos. Pode-se querer tão somente determinar o número de variáveis latentes, isto é, variáveis que não podem ser medidas diretamente mas que estão na origem das observadas, e com isso tentar obter uma representação dos dados em baixa dimensão. Esse é o objetivo dos métodos de redução da dimensionalidade (*Dimensionality Reduction, DR*). No entanto, se além de obter o número de variáveis latentes, também se desejar recuperar tais variáveis, então tal objetivo é chamado de separação das variáveis latentes. A maioria dos métodos não é capaz de cumprir ambos os objetivos, no entanto, pode-se combinar métodos diferentes para alcançar todas as tarefas.

O estudo de dados de alta dimensionalidade possui motivação prática para visualização de dados e análise de dados multivariados, mas também possui motivações teóricas. Com relação à motivação teórica, deve ser considerada a chamada "maldição da dimensionalidade" (*curse of dimensionality*), que envolve o fenômeno do espaço vazio. Esse fato se verifica pois os dados de alta dimensionalidade são esparsos nesse espaço de dimensões elevadas, uma vez que os dados geralmente disponíveis se restringem a algumas observações. Essa constatação resulta em propriedades inesperadas, como a concentração de normas e distâncias, em que o aumento da dimensionalidade diminui o

contraste fornecido pelas métricas usuais. Em Lee et al. (33), podem ser verificadas algumas propriedades inesperadas relacionadas a alta dimensionalidade do espaço.

Alguns exemplos em que dados de alta dimensionalidade são encontrados: processamento de imagens, análise de dados multivariados, mineração de dados, dentre outros. O primeiro passo ao analisar dados de alta dimensionalidade é determinar o número de variáveis latentes. Em seguida, com este número, fornecer uma representação dos dados em baixa dimensão, buscando preservar a estrutura envolvida nos dados originais.

3.1 Dimensionalidade intrínseca dos dados

A maioria dos métodos de redução da dimensionalidade não possui um estimador da dimensionalidade intrínseca dos dados e, portanto, precisam de um algoritmo adicional para estimá-la. Essa estimativa entra no método como um parâmetro externo. Obter tal estimativa deve ser a primeira tarefa de qualquer método de redução da dimensionalidade. A dimensionalidade intrínseca revela a presença de uma estrutura topológica nos dados. De um ponto de vista topológico, determinar o número de variáveis latentes equivale a determinar a dimensionalidade intrínseca dos dados. Assim, o termo dimensionalidade intrínseca e o número de variáveis latentes são usados com o mesmo sentido. As variáveis latentes são também chamadas de graus de liberdade.

A dimensão intrínseca dos dados nos fornece informações quanto a complexidade do sistema que gera os dados, o tipo de modelo necessário para descrevê-lo assim como os graus de liberdade do sistema, os quais, em geral, não são iguais a dimensionalidade do espaço inicial. Quando a dimensão intrínseca k dos dados equivale a dimensão D do espaço de dados original, não existe estrutura, existem graus de liberdade suficientes para que, dado um ponto qualquer, uma bola de raio ϵ centrada nesse ponto, possua muitos outros pontos. Entretanto, se $k < D$, os dados estão frequentemente restritos a uma parte bem delimitada do espaço. Dessa forma, uma dimensão intrínseca baixa indica que o conjunto de dados faz parte de um objeto ou estrutura topológica.

É importante ter uma boa estimativa da dimensionalidade intrínseca dos dados para que se possa descobrir a dimensão da variedade da qual os dados fazem parte e, para realizar um remapeamento dos dados para um novo espaço de dimensão mais baixa. Esse remapeamento deve ocorrer de maneira a preservar a estrutura da variedade.

Dentre os estimadores da dimensionalidade intrínseca há aqueles que usam geometria fractal, tais métodos calculam a chamada dimensão fractal. Há também o método de análise dos componentes principais (*Principal Component Analysis*, PCA), o qual possui um estimador da dimensionalidade intrínseca integrado, baseado no mesmo modelo. Mas, como o método é linear, o estimador só funciona para variedades que possuem dependência linear. Para variedades mais complexas, pode-se utilizar o PCA em uma escala local, isto é, pode-se decompor o espaço em partes menores e considerar cada parte separadamente, assumindo que a variedade é aproximadamente linear em cada uma destas divisões e implementando o PCA. A dimensionalidade é obtida pela média ponderada de todos os PCAs locais. A ponderação é realizada pelo número de pontos em cada parte em que o espaço foi dividido. Ao invés de generalizar o PCA para variedades não-lineares pela divisão do espaço em pedaços menores, podem-se utilizar outros métodos DR que já possuem um modelo não-linear.

Como já mencionado, a maioria dos métodos DR não possuem um estimador da dimensionalidade integrado como o PCA. No entanto, alguns desses métodos minimizam um erro de reconstrução, isto é, erro no mapeamento entre os espaços de alta e baixa dimensionalidade. Neste caso, emprega-se o chamado método da tentativa e erro, no qual a dimensionalidade dos dados originais é reduzida aos poucos e se utiliza validação cruzada, por exemplo, para se obter a dimensionalidade com o menor valor do erro de reconstrução. Para maiores informações sobre estes estimadores, consulte Lee et al. (33).

3.2 Métodos de redução da dimensionalidade

Dentre os métodos de redução de dimensionalidade existentes, existem muitas formas de classificá-los, uma das mais simples reside em diferenciar aqueles que se baseiam em modelos lineares dos que se baseiam em modelos não lineares. Tais modelos são a base dos chamados métodos lineares de redução da dimensionalidade e métodos não lineares de redução da dimensionalidade, respectivamente. Pode-se ainda classificar os métodos de acordo com a utilização de modelos discretos ou modelos contínuos; quanto a possuir a estimativa da dimensionalidade intrínseca integrada ao método ou como um parâmetro externo; com relação ao mapeamento entre os espaços de alta e baixa dimensão ser explícito ou implícito; com relação a efetuar uma otimização aproximada ou exata, dentre outros. Em Lee et al. (33), os diferentes métodos de DR são classificados em um dos dois tipos: aqueles que se baseiam na preservação da distância e aqueles que se baseiam na preservação da to-

pologia. Nesta subseção, faremos uma revisão de alguns métodos DR sem a preocupação de escolher algum critério base para dividi-los.

3.2.1 PCA

A análise dos componentes principais é um dos métodos mais conhecidos e utilizados de redução da dimensionalidade. O método foi primeiro introduzido em 1901 por Pearson (50) e desenvolvido independentemente em 1933 por Hotelling (24). A ideia central do PCA é reduzir a dimensionalidade de um conjunto de dados no qual existe um grande número de variáveis que estão relacionadas entre si, ao mesmo tempo que procura reter ao máximo a variação presente no conjunto de dados. Esta redução é alcançada pela transformação das variáveis em um novo conjunto, as componentes principais, as quais são descorrelacionadas, e ordenadas a fim de que as primeiras componentes principais retenham o máximo da variação presente em todas as variáveis originais. Portanto, as k componentes principais de uma coleção de n variáveis aleatórias ($k < n$), são combinações lineares especias das mesmas e trazem consigo a maior parte da informação contida nas n variáveis originais. O método é simples e, para calcular as componentes principais basta resolver um problema de autovalores e autovetores para uma matriz simétrica definida positiva, a matriz de variância e covariância dos dados. O PCA tem uma enorme aplicabilidade e atinge os seguintes objetivos: estimação da dimensionalidade intrínseca dos dados, redução da dimensionalidade pela projeção, de modo linear, das variáveis observáveis no subespaço latente estimado e separação das variáveis latentes. PCA não é um bom método para trabalhar com conjuntos de dados complexos e, estendê-lo para modelos não lineares é uma alternativa que outros métodos buscam resolver. Para maiores informações sobre o método, consulte Jolliffe (27).

3.2.2 KPCA

O PCA é um método que modela variabilidade linear em dados de alta dimensionalidade. No entanto, muitos conjuntos de dados de alta dimensionalidade têm uma natureza não linear. Nestes casos, os dados podem pertencer ou estar próximos a uma variedade não-linear (não em um subespaço linear) e portanto PCA não pode representar corretamente a variabilidade dos dados. Assim, um dos algoritmos designados para lidar com este problema é o kernel PCA (KPCA). Com o Kernel PCA, através do uso de núcleos, as componentes

principais podem ser calculadas eficientemente em espaços característicos de alta dimensão, que estão relacionados ao espaço de entrada por algum mapeamento não-linear. O kernel PCA não envolve nenhuma otimização não linear. Para maiores detalhes deste método e outros métodos de aprendizagem baseados em núcleos, veja Scholkopf et al. (62).

3.2.3 MDS

O escalonamento multidimensional métrico (*Multidimensional Scaling*, MDS), na sua versão clássica, preserva produtos escalares dois a dois. Embora o método não alcance redução da dimensionalidade de uma forma não linear, é considerado o antecedente de todos os métodos não lineares de preservação das distâncias. O MDS e o PCA minimizam o mesmo critério. Tal equivalência entre o MDS e PCA pode ser vantajosa em algumas situações. Por exemplo, se os dados consistirem de distâncias ou similaridades e as coordenadas dos dados não estiverem disponíveis, pode-se aplicar MDS no lugar do PCA. Por outro lado, mesmo quando se conhecem as coordenadas, um método pode ser computacionalmente menos custoso do que o outro. Por exemplo, se os dados não possuem a dimensionalidade muito alta mas apresentam uma amostra muito grande, o PCA exige menos recursos de memória do que o MDS. Entretanto, o MDS tem um desempenho melhor quando a dimensionalidade é alta mas o número de dados é pequeno. A explicação para estas últimas observações se deve pelo seguinte motivo: considere a matriz Y de dados observados, onde cada coluna representa uma observação, o PCA trabalha com a matriz YY^T e o MDS com a matriz Y^TY, onde a última representação é chamada a matriz Gram. Dada a equivalência entre os métodos, o MDS apresenta as mesmas vantagens e desvantagens que o PCA, ou seja, é simples, robusto, mas estritamente linear. Uma variante do método MDS métrico clássico é o MDS métrico, que preserva distância dois a dois, ao invés do produto escalar. Mais detalhes podem ser obtidos em Lee et al. (33) e Cox et al. (14).

3.2.4 Mapeamento não linear de Sammon

O mapeamento não linear de Sammon (*Sammon's Nonlinear Mapping*) foi proposto em 1969 por Sammon (61). O método proposto é similar ao MDS métrico (veja Lee et al. (33)), e se baseia na minimização da seguinte função custo:

$$\frac{1}{c}\sum_{i=1,i<j}^{N}\frac{(d_{\mathbf{y}}(i,j)-d_{\mathbf{x}}(i,j))^2}{d_{\mathbf{y}}(i,j)},$$

onde:

- $d_{\mathbf{y}}(i,j)$ é a distância entre os pontos i e j no espaço de entrada de dimensão D. Nenhuma hipótese é feita na função distância $d_{\mathbf{y}}(i,j)$, mas geralmente a distância euclidiana é escolhida;
- $d_{\mathbf{x}}(i,j)$ é a distância euclidiana entre os pontos i e j no espaço de dimensão P, $(P < D)$.
- $c = \sum_{i=1,i<j}^{N} d_{\mathbf{y}}(i,j)$

Na função custo de Sammon, pode-se observar que o fator c é inversamente proporcional a distância nos dados de entrada. Dessa forma, a preservação de distâncias longas é menos importante do que a preservação de distâncias mais curtas. Para minimizar a função custo é utilizado uma variante do método de Newton, chamada otimização quase-Newton (Alexey et al. (25), Bertsekas (6). Em comparação com MDS, o método de Sammon pode lidar com variedades não lineares. O método também é muito utilizado para visualização bidimensional de dados multivariados.

3.2.5 LLE

O Mapeamento Topológico Localmente Linear (*Locally Linear Embeding*, LLE) é um método não linear de redução da dimensionalidade apresentado por Roweis e Saul (57, 58), que preserva as relações de vizinhança dos dados de entrada quando mapeados para um espaço de baixa dimensão. Os dados são assumidos pertencerem a uma variedade de baixa dimensão, localmente linear, assim, cada ponto pode ser reconstruído a partir dos seus vizinhos por meio de pesos apropriados. Esses pesos capturam as propriedades geométricas intrínsecas das vizinhanças locais, a saber, aquelas propriedades invariantes a translação, rotação e escalonamento. O algoritmo obteve seu nome devido a natureza das reconstruções, isto é, são locais, já que apenas vizinhos contribuem para cada reconstrução, e linear, já que nas reconstruções são considerados subespaços lineares. A dimensionalidade intrínseca dos dados é um parâmetro externo uma vez que nem sempre é possível estimá-la com os dados. O resultado do LLE também pode ser generalizado para novos locais no espaço de entrada. Por exemplo, pode-se calcular o valor de uma saída $\mathbf{y}$ correspondente a uma nova entrada $\mathbf{x}$. Ao invés de rodar novamente

o algoritmo LLE acrescido do novo dado, pode-se obter um mapeamento explícito entre os espaços de baixa e alta dimensão do LLE, utilizando tanto um modelo paramétrico quanto um não paramétrico, conforme exibido em Roweis (58). LLE pode ser utilizado em problemas não lineares de redução da dimensionalidade, seu procedimento de otimização é simples de implementar e não envolve mínimo local, tendo um custo computacional favorável quando comparado a métodos puramente lineares, como PCA e MDS. O método também gera um problema de autovalores esparso, diferente dos problemas de autovalores densos, encontrados no PCA e MDS.

3.2.6 ISOMAP

O ISOMAP é um método não linear de redução da dimensionalidade, simples, que usa a distância por grafos como uma aproximação para a distância geodésica. O método foi proposto por Tenenbaum et al. em (72) e pode ser visto como uma generalização do MDS. A diferença entre eles é que o ISOMAP utiliza distância por grafos enquanto que o MDS utiliza distância euclidiana. Tal diferença é que torna o método ISOMAP um método não linear. O método encontra o ótimo global da sua função erro, na forma fechada. ISOMAP pode lidar com pontos que não estão no conjunto de dados original realizando uma interpolação. O ISOMAP opera como o MDS métrico, decompondo uma matriz Gram em autovalores e autovetores, sendo frequentemente classificado como um método espectral. Como o ISOMAP posui o mesmo tipo de modelo e procedimento de otimização que o PCA e o MDS métrico, ele também herda a integrada estimação da dimensionalidade intrínseca. Enquanto PCA e MDS são designados apenas para variedades lineares, o ISOMAP pode lidar com variedades não lineares.

3.2.7 Mapeamento não linear geodésico de Sammon

Uma variante do método de Sammon é o Mapeamento não linear geodésico de Sammon, o qual utiliza a distância geodésica no espaço de dados, enquanto mantém a distância Euclidiana no espaço mapeado. A opção em manter a distância euclidiana no espaço mapeado é que ela ajuda a deduzir uma regra de atualização não muito complicada para a otimização da função de custo. Informações sobre essa variante do método de Sammon podem ser encontradas em Estevez et al. (17). Como todos os métodos que usam distância geodésica, o método pode sofrer nos casos em que as distâncias por

grafos fornecem aproximações ruins da distância geodésica.

3.2.8 SDE

O Mapeamento Topológico Semidefinido (*Semidefinite Embeding*, SDE) assim como o MDS métrico, ISOMAP e KPCA implementa a preservação da distância por meio de uma decomposição espectral. No MDS, a distância dois a dois é convertida em produto escalar. No ISOMAP, a distância euclidiana tradicional é substituída por distâncias por grafos. No KPCA, as distâncias euclidianas são não linearmente transformadas por meio de uma função núcleo. Nos métodos mencionados, os resultados dependem de uma transformação específica aplicada as distâncias aos pares, isto é, esta transformação depende de uma escolha a priori: distância euclidiana ou por grafo, kernel polinomial ou gaussiano. No entanto, a ideia do SDE é determinar essa transformação de maneira que a mesma seja fornecida pelos dados. Para este fim, as distâncias estão restritas a serem preservadas apenas localmente. A relaxação quanto a preservação estrita da distância, para uma condição mais amena de isometria local, permite ao SDE mapear variedades de forma não linear. O método foi proposto por Weinberger e Saul (47). Para obter SDE, a ideia é aprender a matriz núcleo como um exemplo de programação semidefinida. Como a matriz núcleo representa produtos internos de vetores em um espaço de Hilbert, ela deve ser semidefinida positiva. O núcleo também deve ser centrado e deve-se impor restrições na matriz núcleo para garantir que as distâncias e ângulos entre pontos e seus vizinhos sejam preservadas sobre um grafo da vizinhança, isto é, se dois dados são vizinhos ou vizinhos comuns de outra entrada, então a distância deve ser preservada. O SDE constrói um programa semidefinido para aprender a matriz núcleo. O LLE propõe ainda outra aproximação baseada em mapeamentos que preservam ângulos locais. A preservação de ângulos e distâncias locais estão de certa forma relacionadas e podem ser interpretadas como duas maneiras diferentes de preservar produtos escalares locais.

3.2.9 Automapa Laplaciano

O Automapa Laplaciano (*Laplacian Eigenmaps*) é similar ao LLE e foi desenvolvido em 2003 por Belkin e Niyogi (4). Dado um conjunto de pontos em um espaço, constrói-se um grafo ponderado, com arestas conectando pontos próximos. O grafo da vizinhança pode ser obtido encontrando-se os k vizinhos mais próximos ou considerando todos os pontos dentro de um raio

fixo ϵ. Para ponderar as arestas, pode-se utilizar um dos seguintes pesos: $W_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{s}}$, onde s é um parâmetro escolhido a priori, ou utiliza-se $W_{ij} = 1$,se os vértices i e j estão conectados. Um mapeamento é realizado através do cálculo de autovetores do grafo laplaciano. O Laplaciano de um grafo é análogo ao operador Laplace Beltrami em variedades. A aproximação utiliza as propriedades do operador Laplace Beltrami para construir mapas invariantes para a variedade. Enquanto tais mapas têm algumas propriedades preservando localidade, eles não fornecem, em geral, um mapeamento isométrico.

3.2.10 HLLE

Donoho e Grimes (16) propuseram o Hessiano LLE (*Hessian-based Locally Linear Embedding*, HLLE). O método busca recuperar uma parametrização dos dados pertencentes a uma variedade M, que é localmente isométrica a um subconjunto conexo, aberto do espaço euclidiano. O método pode ser visto como uma modificação do LLE e, sua estrutura teórica como uma modificação da estrutura dos automapas laplacianos, utilizando uma forma quadrática baseada na hessiana no lugar de uma baseada no Laplaciano. Para definir a hessiana, são utilizadas coordenadas ortogonais nos planos tangentes de M.

3.2.11 SOM

Os Mapas Auto-Organizáveis (*Self-Organizing Maps*, SOM) juntamente com o Perceptron de múltiplas camadas (*Multilayer Perceptron*) é talvez o método mais amplamente conhecido no campo de Redes Neurais Artificiais. O algoritmo é também chamado Mapa Auto-Organizável de Konohen, pois foi introduzido por kohonen (29) no campo de Redes Neurais. O método realiza simultaneamente a combinação de duas tarefas: a quantização vetorial e redução da dimensionalidade. A quantização vetorial substitui os dados originais por um conjunto menor de pontos, chamados protótipos, e os mesmos devem ser representativos dos dados originais que substituem. O SOM é um método de redução da dimensionalidade que busca a preservação da topologia. O SOM utiliza um modelo de mapeamento discreto e a topologia é descrita de uma forma discreta também. A representação discreta da topologia é geralmente chamada de reticulado, que pode ser um conjunto de pontos regularmente espaçados no plano ou um grafo. No último caso, pontos estão associados com vértices do grafo e a proximidade entre eles é simbolizada por

uma aresta (ponderada) no grafo. No SOM, a redução da dimensionalidade se deve ao uso de um reticulado pré definido e na maioria das implementações a escolha é por uma rede retangular ou hexagonal, com pontos regularmente espaçados. Considerando o algoritmo, o método é simples e é frequentemente utilizado para visualizar dados rotulados (Lee et al. (33)).

3.2.12 GTM

Uma alternativa ao SOM é o Mapeamento Topográfico Generativo (*Generative Topographic Mapping*, GTM)(Bishop et al. (8), Svensen (68)). Em modelagem generativa, a todas as variáveis do problema são atribuídas uma distribuição de probabilidade, no qual a teoria Bayesiana é aplicada para aprendizagem dos dados. Na aprendizagem tradicional ou frequentista, não se assume nenhuma distribuição sobre os parâmetros do modelo. Ao contrário, na aprendizagem bayesiana, uma distribuição de probabilidade sobre os parâmetros do modelo é obtida antes de qualquer dado ser considerado. Esta distribuição é baseada em uma distribuição a priori dos parâmetros, que expressa uma crença inicial sobre o valor dos parâmetros. Em seguida, fornecido o dado, a distribuição a priori é atualizada para uma distribuição a posteriori usando a regra de Bayes. Esta aproximação probabilística utiliza a técnica de otimização dada pelo algoritmo EM. Com relação ao algoritmo, enquanto o GTM otimiza uma função objetivo bem definida, a saber a log verossimilhança, o SOM não possui nenhuma função objetivo explícita ou critério de erro a ser otimizado. Assim como o SOM, o GTM também é limitado a espaços latentes de baixa dimensão, usualmente redes de uma ou duas dimensões.

3.2.13 ISOTOP

O ISOTOP (Lee et al. (32)) é um método que visa superar algumas das limitações do SOM, quando este é usado para redução não linear da dimensionalidade. ISOTOP separa a quantização vetorial (que se torna opcional) e a redução da dimensionalidade. O problema se divide em três passos: quantização vetorial (opcional), construção de grafo e mapeamento em dimensão baixa. O método assume a hipótese de que os pontos pertencem ou estão próximos a uma variedade diferenciável. O primeiro passo é opcional, assim, a quantização vetorial é realizada caso o conjunto de dados possua muitos pontos e, assim, reduzir este número. No entanto, nenhuma relação de vizinhança, ao contrário do SOM, é considerada entre os protótipos. No segundo passo, caso

a quantização vetorial seja realizada, o ISOTOP relaciona pontos de dados vizinhos ou protótipos, usando regras de construção de grafos. Tipicamente, cada ponto está associado a um vértice do grafo e conectado com seus k vizinhos mais próximos ou com todos os pontos dentro de uma bola de raio ϵ. A estrutura do grafo tenta capturar as relações de vizinhança da variedade com base nos dados. O segundo passo termina com o cálculo da distância entre todos os pares de vértices no grafo. Este segundo passo nos fornece um conjunto de pontos conectados, comparáveis ao reticulado retangular de um SOM, exceto que a forma da estrutura é dada pelos dados. No terceiro passo, as coordenadas de alta dimensão de cada dado ou protótipo são substituídas por outras de baixa dimensão, inicialmente nulas. É implementada uma regra de aprendizagem que transforma a estrutura conectada em um espaço de baixa dimensão, buscando preservar as vizinhanças. Assim como o SOM, ISOTOP utiliza um modelo não linear e faz parte da teoria de Redes Neurais Artificiais e usa técnicas aproximadas de otimização. O mapeamento dado pelo ISOTOP, entre os espaços de alta e baixa dimensão, é discreto e explícito, tornando a generalização para novos pontos uma tarefa difícil.

3.2.14 LTSA

O método Alinhamento do Espaço Tangente Local (*Local Tangent Space Alignment*, LTSA) (Zhang et al. (81)) é um algoritmo para aprendizagem de variedades e redução não linear da dimensionalidade. A geometria local da variedade é aprendida pela construção de um espaço tangente local para cada ponto, e, aqueles subespaços tangentes são então alinhados para fornecer as coordenadas globais internas dos dados relativas a variedade.

3.3 Aproximação de Funções

Considere os pares de dados $\mathbf{x}_i \in \mathbb{R}^n, y_i \in \mathbb{R}$ como pontos em um espaço euclidiano de dimensão $(n+1)$. A função $f(\mathbf{x})$ tem domínio igual ao subespaço de entrada de dimensão n e se relaciona com os dados via um modelo tal como $y_i = f(\mathbf{x}_i) + \epsilon_i$. Pode-se assumir, por conveniência, que o domínio é $\mathbb{R}^n$, um espaço euclidiano de dimensão n. Caso a função possua múltiplas saídas, pode-se considerar o problema como sendo um de múltiplas funções com uma única saída. O objetivo consiste em obter uma aproximação útil para $f(\mathbf{x})$, para todo $\mathbf{x}$, em alguma região de $\mathbb{R}^n$. Tratar o problema de aprendizagem supervisionada como um problema de aproximação

de funções permite que os conceitos geométricos do espaço euclidiano e conceitos matemáticos de inferência estatística sejam aplicados ao problema (Trevor et al. (23)). Revisaremos alguns métodos de aproximação de funções que trabalham com grandes conjuntos de dados em altas dimensões.

3.3.1 Redes Neurais

Redes Neurais são frequentemente empregadas como métodos para aproximar funções. O problema de aprender um mapeamento dos dados de entrada/saída a partir de um conjunto de exemplos, realizado por muitas redes neurais, é considerado por Poggio e Girosi (53) como uma aproximação de uma função multidimensional que resolve o problema de reconstrução de uma hipersuperfície[1]. Eles consideraram o problema da aproximação de mapeamentos não lineares, principalmente contínuos. A aproximação é baseada em técnicas de regularização que levam a uma classe de redes de três camadas chamadas redes de regularização e incluem um caso especial do método de funções de base radial. No estudo, a teoria de redes de regularização é generalizada para uma formulação que inclui *clustering* e redução da dimensionalidade. Existem muitas pesquisas baseadas em redes neurais, Ferrari e Stengel (18), por exemplo, utilizam aproximação algébrica para representar funções não lineares, multidimensionais, por rede neurais com fluxo de dados para frente (*feed-forward*). O conjunto de treinamento é associado aos parâmetros ajustáveis da rede por equações peso não lineares. A estrutura em cascata dessas equações permite que elas sejam tratadas como conjuntos de equações lineares. Assim, o processo de treinamento e as propriedades de aproximação da rede podem ser investigados via álgebra linear. São desenvolvidos quatro algoritmos baseados na solução exata ou aproximada das equações peso gradiente e pares de entrada/saída. As implementações realizadas mostraram que treinamento de redes neurais algébricas é rápido e simples para aproximação de funções não lineares sem ruído, além de apresentarem melhores propriedades de generalização do que as técnicas de otimização contemporâneas.

3.3.2 Máquina de Suporte Vetorial

A Máquina de Suporte Vetorial (*Support Vector Machine*, SVM) é um método de aprendizagem supervisionada para classificação, originalmente

[1]Hipersuperfície é uma variedade de dimensão $(N-1)$ no espaço $\mathbb{R}^N$. Uma hipersuperfície é portanto o conjunto de soluções de uma única equação $f(x_1, \ldots, x_N) = 0$ e, tem codimensão um.

desenvolvido por Vapnik et al. (78). Este método foi estendido para regressão, sendo chamado máquina de suporte vetorial para regressão (SVR). A ideia do método é mapear os vetores do espaço de entrada para um espaço característico de alta dimensionalidade utilizando uma função não linear e, nesse espaço efetuar uma regressão linear nos dados. O método utiliza o chamado "*kernel trick*", o qual permite efetuar cálculos implicitamente no espaço de alta dimensão. Treinar o SVR equivale a resolver um problema de programação quadrática convexa. Ademais, o método possui uma representação esparsa da solução, pois a solução do SVR depende apenas de um subconjunto dos dados de entrada, chamados de vetores de suporte.

3.3.3 Processo Gaussiano

Um Processo Gaussiano (*Gaussian Process*, GPs) é uma generalização da distribuição de probabilidade gaussiana. Uma alternativa poderosa para aproximação de funções em espaço de alta dimensionalidade é o Processo Gaussiano para Regressão (GPR) (Ramussen et al. (54)). Dado um conjunto de n pontos de dados treino $\{\mathbf{x}_i, y_i\}_{i=1}^n$, gostaríamos de aprender uma função $f(\mathbf{x}_i)$ transformando o vetor de entrada $\mathbf{x}_i$ no seu valor alvo y_i dado um modelo $y_i = f(\mathbf{x}_i) + \epsilon_i$, onde ϵ_i é um ruído gaussiano com média zero e variância σ^2. Os alvos observados podem ser descritos por uma distribuição gaussiana $y \sim \mathcal{N}(\mathbf{0}, \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I})$, onde $\mathbf{X}$ denota o conjunto contendo todos os dados de entrada $\mathbf{x}_i$ e $\mathbf{K}(\mathbf{X}, \mathbf{X})$ a matriz de covariância calculada usando uma função de covariância dada. Núcleos gaussianos são as funções covariância mais utilizadas. A distribuição conjunta dos valores alvos observados e de um valor previsto $f(\mathbf{x}_*)$ para algum dado $\mathbf{x}_*$ é também uma distribuição normal. A distribuição conjunta fornece o valor médio previsto $f(\mathbf{x}_*)$ e sua variância $V(\mathbf{x}_*)$. O valor ótimo para um conjunto de dados particular pode ser automaticamente estimado pela maximização da log verossimilhança marginal usando métodos de otimização padrão tais como métodos Quase-Newton. Em Nguyen-Tuong et al. (48) é proposto uma aproximação local para o GPR padrão.

3.3.4 Regressão Bayesiana

A técnica de Regressão Bayesiana (*Bayesian Regression*) pode ser usada para incluir parâmetros de regularização na regressão linear, os quais devem ser ajustados aos dados. A regularização l_2 usada na *Ridge Regression* equivale

a encontrar uma solução máxima a posteriori sobre uma priori Gaussiana nos parâmetros ω com precisão $\lambda - 1$. Para obter um modelo probabilístico completo, a saída y é assumida ter uma distribuição gaussiana em torno de $X\omega = \omega_0 + \omega_1 x_1 + \ldots + \omega_p x_p$ onde $\omega = (\omega_1, \ldots, \omega_p)$ e ω_0 é o intercepto:

$$p(y|X, \omega, \alpha) = \mathcal{N}(y|X\omega, \alpha)$$

O método *Bayesian Ridge Regression* estima um modelo probabilístico da regressão, onde a priori para o parâmetro ω é dada por uma gaussiana esférica:

$$p(\omega|\lambda) = \mathcal{N}(\omega|0, \lambda^{-1}\mathbf{I_p})$$

As prioris sobre α e λ são escolhidas serem distribuições gama, o conjugado a priori para a precisão da Gaussiana. O modelo resultante é chamado *Bayesian Ridge Regression*. Os parâmetros ω, α e λ são estimados conjuntamente durante o ajuste do modelo. Os hiperparâmetros restantes são os parâmetros das prioris gama sobre α e λ. Uma boa introdução aos métodos bayesianos pode ser obtida em Bishop (7).

3.3.5
Estimativa de Máximo a Posteriori de Processos Gaussianos

Smola e Bartlett (66) desenvolveram uma técnica simples, ágil e esparsa para aproximar a Estimativa de Máximo a Posteriori (MAP) de Processos Gaussianos, expandindo ela em termos de um subconjunto pequeno de funções núcleos. Em resumo, dado um conjunto de funções núcleo, é procurada uma função adicional que aumente ao máximo a probabilidade a posteriori. Tal função é adicionada ao conjunto de funções base e o processo é repetido até que o máximo seja satisfatoriamente aproximado. Uma aproximação similar para uma cota superior precisa, na probabilidade posterior, dá um critério de parada.

3.3.6
Máquina de Conselho Bayesiano

A Máquina de Conselho Bayesiano (*Bayesian Committee Machine*, BCM) (Tresp (77)) é introduzida como uma solução aproximada para regressão, cujo custo computacional apenas aumenta linearmente com o número de padrões de treino e é aplicável, em particular, para sistemas de regressão baseados em núcleos, porém, também se aplica a outros modelos de regressão. A ideia do método é dividir o conjunto de dados em M conjuntos de dados e treinar

M sistemas nestes conjuntos de dados. No fim, as previsões são combinadas usando um novo esquema de pesos na forma de uma BCM. Uma aplicação é feita na aprendizagem online onde os dados chegam em sequência e o treinamento deve ser realizado em sequência também. O método é aplicado a sistemas de regressão com núcleos, em particular, GPR. O método é ainda aplicado para regressão com funções base fixas.

3.3.7 SRM, RRA e BCM

Tresp e Schwaighofer (63) compararam experimentalmente três aproximações principais voltadas para escalonar GPR para grandes conjuntos de dados: o subconjunto de representantes do método (*Subset of Representers Method*, SRM), a aproximação de posto reduzido (*Reduced Rank Aproximation*, RRA) e o BCM. A diferença entre os métodos é analisada, tanto do ponto de vista teórico quanto do ponto de vista experimental. Uma das maiores diferenças discutidas é que o BCM realiza *transduction*, isto é, explora o conhecimento sobre a localização do dado teste em sua aproximação. Dessa forma, a aproximação BCM é calculada quando as entradas para os dados testes são conhecidas. Por outro lado, os métodos RRA e SRM realizam o estilo de aprendizagem por indução, que significa que os parâmetros do modelo são calculados apenas baseados nos dados de treinamento.

3.3.8 SVR Bayesiano

Chu et al. em (13) usaram uma função de perda para SVR Bayesiano. Foi utilizado o GPR para configurar a estrutura bayesiana, na qual a função de perda definida é utilizada no cálculo da verossimilhança. Com esta estrutura, a estimativa MAP dos valores da função correspondem a solução de um problema SVR estendido. A aproximação global tem os méritos do SVR e dos métodos bayesianos.

3.3.9 RVM

Tipping (74) introduziu uma estrutura Bayesiana geral para obter soluções esparsas para tarefas de regressão e classificação utilizando modelos lineares nos parâmetros. A aproximação apresenta uma especialização denotada a "máquina de vetores de relevância" (*Relevance Vector Machine*, RVM), um modelo de forma funcional idêntica ao SVM. Com a estrutura de aprendizagem

Bayesiana, são obtidos modelos de previsão precisos os quais utilizam menos funções base do que um SVM comparável, além de oferecer vantagens tais como o benefício de previsões probabilísticas e a facilidade de utilizar funções base arbitrárias. É introduzido um modelo de aprendizagem Bayesiana Esparso para Regressão para o RVM.

3.3.10 RCVI

Nesta tese, a partir da aprendizagem da variedade utilizando votação por tensores, obtemos uma estimativa da dimensionalidade intrínseca dos dados bem como informações confiáveis das tangentes e normais em cada ponto. Com estas informações, buscamos aproximar variedades, considerando-as como o conjunto de nível zero de uma função polinomial multivariada. No entanto, para obter a função, é realizada uma partição da unidade, isto é, é realizada uma subdivisão do domínio, controlada pelo erro, e os dados são aproximados em cada subdomínio separadamente por uma função polinomial multivariada e, por fim, as soluções locais são unidas usando pesos locais que somam para um em toda parte do domínio. Essa partição do domínio controlada pelo erro dá a estrutura de uma árvore. Após a obtenção dessa árvore, podemos realizar uma regressão.