Notação
Negrito = vetor.
Exemplos:
$\mathbf{x}$ é um vetor $[x_0, ..., x_n]$
$\boldsymbol{\phi}(x)$ é um vetor de funções $[\phi_0(x), ..., \phi_n(x)]$

1 Regressão

O objetivo da regressão é fazer previsões dos valores de uma *target variable* t dado o valor de uma *input variable* $\mathbf{x}$, $\mathbf{x}$ é um vetor de dimensão qualquer.

Dado um conjunto de dados composto de observações $\{\mathbf{x}_n\}$ e suas respectivas *target variables* $\{t_n\}$, o que se quer fazer é a previsão do valor de t dado um valor de $\mathbf{x}$ que não necessariamente está presente nas observações. Em outras palavras, quer-se encontrar uma função $y(\mathbf{x})$ que associa cada valor possível de $\mathbf{x}$ a uma previsão de t.

A função $y(\mathbf{x})$ pode ser encontrada através do ajuste de parâmetros. Considere $\mathbf{w}$ como um vetor de parâmetros ajustáveis; os valores finais dos parâmetros ajustáveis $\mathbf{w}_f$ devem ser escolhidos de maneira que, a função $y(\mathbf{x}) = y(\mathbf{x}, \mathbf{w}_f)$, minimize (ou quase isso) o valor de $\mathcal{L}(y)$, onde $\mathcal{L}$ é uma função de perda adequada. A função de perda $\mathcal{L}$ é o criterio pelo qual se julga o quão adequadas são as previsões de y dado o conjunto de dados.

O método pelo qual os parâmetros ajustáveis são atualizados aqui é o da *descida de gradiente*, onde os parâmetros são modificados de maneira iterativa na direção do negativo do gradiente da função de perda,

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \lambda \nabla \mathcal{L}(y)$$

ou seja, a cada passo os parâmetros ajustáveis $\mathbf{w}$ são modificados de maneira a diminuir o valor da função de perda $\mathcal{L}(y)$.

2 Modelos Lineares de Regressão

Os modelos lineares de regressão são aqueles cuja função $y(\mathbf{x}, \mathbf{w})$ é linear em relação aos parâmetros ajustáveis $\mathbf{w}$, o que não quer dizer que $y(\mathbf{x}, \mathbf{w})$ seja necessariamente linear com relação à $\mathbf{x}$. No geral, $y(\mathbf{x}, \mathbf{w})$ é uma combinação linear de *basis functions* $\phi(\mathbf{x})$, funções de $\mathbf{x}$ que podem ou não ser lineares,

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{M-1} w_i \phi_i(\mathbf{x}_n)$$

onde M é a quantidade de parâmetros ajustáveis. A equação anterior pode ser simplificada para $\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$ usando notação vetorial, assumindo que $\phi_0 = 1$.

A função de perda utilizada aqui é a soma dos quadrados dos erros, isto é, as diferenças entre as previsões do modelo e o valor real das observações $t_n - y(\mathbf{x}_n, \mathbf{w})$ são elevadas ao quadrado e somadas, por todas as instâncias de observação, para gerar um valor que representa o quão adequada é a escolha de valores para $\mathbf{w}$.

$$\mathcal{L}(y(\mathbf{w})) = 1/2\sum_{n=1}^{N}\{t_n - y(\mathbf{x}_n, \mathbf{w})\}^2$$

onde N é a quantidade de observações. Tal escolha de função de perda é justificada se assumirmos que os dados observados são gerados a partir de uma função determinística somada a um ruído gaussiano.

Essa escolha de função de perda, combinada ao fato de $y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})$ ser linear em relação à $\mathbf{w}$, nos leva à sequinte equação para o gradiente da função de perda

$$\nabla\mathcal{L}(y(\mathbf{w})) = \sum_{n=1}^{N}\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\}\boldsymbol{\phi}(\mathbf{x}_n)^T$$

.

3 Referências

Pattern Recognition and Machine Learning - Christopher Bishop, Capítulo 3.