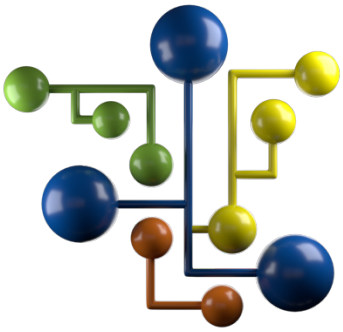


Data Science Academy

www.datascienceacademy.com.br

Introdução à Inteligência Artificial

Regressão com Modelos Lineares Multivariados

Podemos facilmente estender de regressão linear simples, para problemas de regressão linear multivariada (ou múltipla), em que cada exemplo x_j é um vetor de n elementos. Nosso espaço de hipótese é o conjunto de funções da forma:

$$h_{sw}(\mathbf{x}_j) = w_0 + w_1 x_{j,1} + \cdots + w_n x_{j,n} = w_0 + \sum_i w_i x_{j,i}$$

O termo w_0, a interseção, distingue-se como diferente dos outros. Podemos corrigir isso pela criação de um atributo de entrada fictício, $x_{j,0}$, que sempre é definido como igual a 1. Então, h é simplesmente o produto escalar dos pesos e do vetor de entrada (ou, de forma equivalente, o produto da matriz transposta dos pesos pelo vetor de entrada):

$$h_{sw}(\mathbf{x}_j) = \mathbf{w} \cdot \mathbf{x}_j = \mathbf{w}^\top \mathbf{x}_j = \sum_i w_i x_{j,i}$$

O melhor vetor de pesos, w*, minimiza a perda de erro quadrático nos exemplos:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_j L_2(y_j, \mathbf{w} \cdot \mathbf{x}_j)$$

A regressão linear multivariada não é muito mais complicada do que o caso univariado já abordado. A descida pelo gradiente vai atingir o mínimo (único) da função de perda; a equação de atualização de cada peso w_i é:

$$w_i \leftarrow w_i + \alpha \sum_j x_{j,i}(y_j - h_{\mathbf{w}}(\mathbf{x}_j))$$

Também é possível resolver analiticamente para o w que minimiza a perda. Seja y o vetor de saída para os exemplos de treinamento e X a matriz de dados, ou seja, a matriz de entradas com um exemplo n dimensional por linha. Então, a solução

$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

minimiza o erro quadrático.

Com a regressão linear univariada não precisamos nos preocupar com a superadaptação (overfitting). Mas, com a regressão linear multivariada em espaços de dimensão superior, é possível que alguma dimensão que seja realmente irrelevante pareça ser útil por acaso, resultando em superadaptação.

Assim, é comum o uso de regularização em funções lineares multivariadas para evitar a superadaptação. Lembre-se de que, com a regularização, minimizamos o custo total de uma hipótese, contando tanto com a perda empírica como com a complexidade da hipótese:

$$Custo(h) = PerdaEmp(h) + \lambda\ Complexidade(h)$$

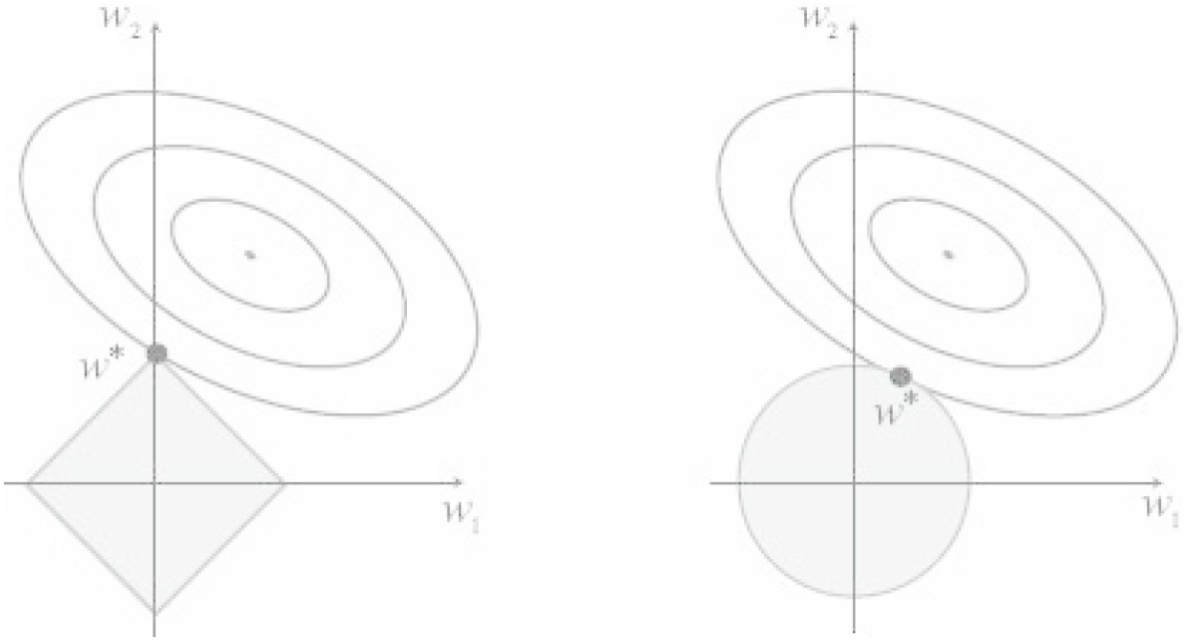
Para as funções lineares, a complexidade pode ser especificada em função dos pesos. Podemos considerar uma família de funções de regularização:

$$Complexidade(h_{\mathbf{w}}) = L_q(\mathbf{w}) = \sum_i |w_i|^q$$

Tal como acontece com as funções de perda com θ = 1 temos a regularização L_1, o que minimiza a soma dos valores absolutos; com θ = 2, a regularização L_2 minimiza a soma dos quadrados. Qual função de regularização se deve escolher? Isso depende do problema específico, mas a regularização L_1 tem uma vantagem importante: tende a produzir um modelo esparso. Isto é, muitas vezes define muitos pesos para zero, declarando efetivamente os atributos correspondentes como irrelevantes — como a APRENDIZAGEM-EM-ÁRVORE-DE-DECISÃO faz (embora por um mecanismo diferente). As hipóteses que descartam atributos podem ser mais fáceis de um ser humano entender, e podem ser menos prováveis de superadaptar.

A figura abaixo fornece uma explicação intuitiva da razão pela qual a regularização L_1 leva a pesos zero, enquanto a regularização L_2 não leva. Observe que a minimização [Perdas(w) + λ Complexidade(w)] é equivalente à minimização Perda(w) sujeita à restrição de que Complexidade(w) ≤ c, para alguma constante c que esteja relacionada com λ. Agora, na figura abaixo (do lado esquerdo) a caixa em forma de losango representa o conjunto de pontos w no espaço de peso bidimensional que tem a complexidade L_1 menor que c; nossa solução terá de estar em algum lugar dentro dessa caixa. As ovais concêntricas representam contornos da função de perda, com a perda mínima ao centro. Queremos encontrar o ponto na caixa que esteja mais próximo ao mínimo; você pode observar no diagrama que, para uma posição arbitrária de mínimo e seus contornos, será comum para o canto da caixa encontrar sua forma mais próxima ao mínimo só porque os cantos são pontudos. E, claro, os cantos são os pontos que têm valor de zero em alguma dimensão. Na figura abaixo (do lado direito), fizemos o mesmo para a medida de complexidade L_2, que representa um círculo, em vez de um diamante. Aqui você pode verificar que, em geral, não há razão para a interseção aparecer em um dos eixos; assim, a regularização L_2 não tende a produzir pesos zero. O resultado é que o número de exemplos necessários para encontrar um bom h é linear

no número de características irrelevantes para a regularização L_2, mas somente com a regularização logarítmica L_1. A evidência empírica de muitos problemas reforça essa análise.



Porque a regularização de L_1 tende a produzir um modelo esparso. (a) Com a regularização (caixa) de L_1, a perda mínima atingível (contornos concêntricos), muitas vezes, ocorre em um eixo, o que significa peso zero. (b) Com a regularização L_2 (círculo), é provável que a perda mínima ocorra em qualquer parte do círculo, não dando preferência a pesos zero.

Outra maneira de ver isso é que a regularização L_1 leva os eixos dimensionais a sério, enquanto a L_2 trata-os como arbitrários. A função de L_2 é esférica, o que a torna rotacionalmente invariante: imagine um conjunto de pontos em um plano, medido por suas coordenadas x e y. Agora imagine a rotação de 45º dos eixos. Você gostaria de obter um conjunto de valores diferentes de (x', y') representando os mesmos pontos. Se aplicar a regularização L_2 antes e depois da rotação, terá exatamente o mesmo ponto como resposta (embora o ponto seja descrito com a nova coordenada (x', y')). Isso é apropriado quando a escolha dos eixos realmente for arbitrária — quando não importar se as suas duas dimensões são as distâncias norte e leste ou nordeste e sudeste. Com a regularização L_1 obtém-se uma resposta diferente porque a função L_1 não é rotacionalmente invariante. Isso é apropriado quando os eixos não são intercambiáveis.

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig