A1: MACHINE LEARNING

REGRESSÃO LINEAR

Um conjunto de dados $\mathcal{D}$ com N pares ordenados (x_n, y_n). Os $x_n \in \mathbb{R}^D$ são variáveis independentes e y_n são amostras de uma variável dependente.

Queremos o vetor de pesos θ que minimiza a perda quadrática, $\hat{\theta}_{LS} = \arg\min_{\theta \in \mathbb{R}^{D+1}} \{l(\theta) := \frac{1}{N} \sum_{i=n}^{N} (y_n - \theta^T x_n)^2\}$.

$\hat{\theta}_{LS} = (X^T X)^{-1} X^T y$ no caso em que $X^T X$ é definida positiva.

Se $N > D$ e $X^T X$ não é definida positiva, utilizamos a inversa de $X^T X + \alpha I$.

Se $N = D$ e a matriz X possui inversa, $\hat{\theta}_{LS} = X^{-1} y$.

Se $N < D$, daí $\hat{\theta}_{MN} = X^T (X X^T)^{-1} y$.

Aqui, $\hat{\theta}_{LS} = \hat{\theta}_{EMV}$ (derivamos log-verossimilhança da Gaussiana e igualamos a 0). Minimizar o MSE é equivalente a admitir uma verossimilhança da forma $y|x \sim \mathcal{N}(\theta^T x, \sigma^2)$.

Regressões lineares **parametrizam a média** como uma combinação linear dos vetores de entrada.

EXPANSÃO DE BASE: a relação entre entrada e saída não é necessariamente linear. Usamos uma transformação não-linear das variáveis de entrada: $\hat{y} = \theta^T \phi(x)$. O processo a partir daí é o mesmo. Alguns exemplos são:

Polinômios: $\phi(x) = [x^2, x, 1]$, $\phi(x) = [x_1^2, x_2^2, x_1 x_2, x_2, 1]$;

Funções de base radiais: dados centros (pontos) $c_1, c_2, ..., c_M$, $\phi(x) = [f(||x - c_1||), ..., f(||x - c_M||)]^T$. Classe de redes neurais chamada redes RBF. Para aplicar, basta determinar os centros (em geral usando *KMeans*) e a função f_{c_i}, normalmente:

- a Gaussiana, $f_{c_i}(x) = \exp(-\gamma ||x - c_i||_2^2)$,
- ou a Multi-quadrática, $f_{c_i}(x) = \sqrt{1 + \epsilon ||x - c_i||_2^2}$

OTIMIZAÇÃO

Queremos otimizar o vetor $\theta \in \mathbb{R}^m$ utilizando abordagens restritas (todo $\mathbb{R}^m$) ou irrestritas (subconjunto do $\mathbb{R}^m$).

Otimização irrestrita: queremos θ que $\min_{\theta \in \mathbb{R}^m} l(\theta)$. θ^* é a solução ótima, mas não necessariamente existe (quando $l(\theta)$ não possui limite inferior, e.g.). Condição necessária $\nabla_\theta l(\theta^*) = 0$.

Se não há solução analítica: partimos de $\theta^{(0)}$ e a cada iteração $t = 1, 2, ..., T$, queremos $\theta^{(t)}$ tal que $l(\theta^{(t)}) < l(\theta^{(t-1)})$. Utilizamos o resultado como aproximação para θ^*.

Gradiente Descendente: damos passos na direção oposta ao gradiente da função custo à cada iteração. A partir de $\theta^{(0)}$, iteramos $\theta^{(t+1)} = \theta^{(t)} - \alpha^{(t)} \nabla l(\theta^{(t)})$. Utilizamos o mesmo α em todas as iterações.

Gradiente Descendente Estocástico: Para N grande, é custoso minimizar $f(\theta) = \sum_{n=1}^{N} f_n(\theta)$. Amostramos os dados aleatoriamente para atualizar os parâmetros.

Random Shuffling: permuta os índices e processa os dados sequencialmente sequencialmente, usando M termos para cada atualização do gradiente.

Multiplicadores de KKT: $\mathcal{L}(\theta, \mu, \lambda) := l(\theta) + \mu^T g(\theta) + \lambda^T h(\theta)$, onde a função $l(\theta)$ deve ser minimizada sujeita às restrições das funções $g(\theta)$ e $h(\theta)$.

REGRESSÃO LOGÍSTICA

Abordagem frequentista: y tem distribuição é Bernoulli com parâmetro $g(x)$. g mapeia o valor resultante para o intervalo [0, 1] (*logit*). Utilizamos a sigmoide $\sigma(t) = (1 - \exp(-1))^{-1}$ para definir:

$$p(y|x) = \text{Ber}(y|\sigma(\theta^T x)) = \sigma(\theta^T x)^y (1 - \sigma(\theta^T x))^{1-y}$$

Sendo a verossimilhança $\mathcal{L}(\theta) = \prod_{i=1}^{N} \sigma(\theta^T x_i)_i^y (1 - \sigma(\theta^T x_i))^{1-y_i}$, então $\hat{\theta} = \arg\max_{\theta \in \mathbb{R}^{D+1}} \mathcal{L}(\theta) = \arg\min_{\theta \in \mathbb{R}^{D+1}} -\log \mathcal{L}(\theta)$.

$-\log \mathcal{L}(\theta)$ é convexa, mas sua minimização não possui solução analítica. Encontramos $\hat{\theta}$ por otimização.

TEORIA DA INFORMAÇÃO: interpreta-se $\log 1/q(z)$ como uma medida de *surpresa*, i.e., do quanto observar um valor específico z contrasta com seu conhecimento prévio, representado por q. $H(p, q)$ é o valor dessa medida se os valores de z são amostrados de p ao invés de q.

$$H(p, q) = \mathbf{E}_{z \sim p} \left[\log \frac{1}{q(z)} \right]$$

$q = p$ minimiza H, e aí $H(p) := H(p, p)$ se chama entropia. A entropia é máxima quando p é uma distribuição uniforme (todo valor tem a mesma chance de ser observado).

O logaritmo negativo da verossimilhança Bernoulli é a entropia cruzada binária. Para $-\log \text{Ber}(y|r)$, define-se $p(z) = \text{Ber}(z|y)$ e $q(z) = \text{Ber}(z|r)$.

Daí $H(p, q) = -y \log q(1) - (1 - y) \log q(0) = -(y \log r + (1 - y) \log(1 - r))$.

Isso implica $\exp(-H(p, q)) = r^y (1 - r)^{(1-y)} = \text{Ber}(y|r)$.

Abordagem bayesiana: reflete incerteza sobre o valor estimado. Dada uma *priori* $p(\theta)$, computamos a distribuição de θ condicionada nos dados $\mathcal{D}$, a *posteriori*. Representa a incerteza sobre o valor de θ:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) p(\theta)}{\int_\theta p(\mathcal{D}|\theta) p(\theta)}$$

Procura-se o ponto que maximiza a posteriori e toma-se como estimativa pontual.

Máximo *a posteriori*: $\hat{\theta}_{\text{MAP}} = \arg\max_\theta \log p(\mathcal{D}|\theta) + \log p(\theta)$. Versão regularizada do EMV, porque $\log p(\theta)$ penaliza regiões pouco prováveis a priori.

Aproximação de Laplace: aproximar $p(\theta|\mathcal{D})$ por $q(\theta)$ usando uma expansão de Taylor de segunda ordem em $\log p(\theta|\mathcal{D})$ ao redor da moda da posteriori ($\hat{\theta}_{\text{MAP}}$). $H = \nabla_\theta^2 - \log p(\theta|\mathcal{D})|_{\theta=m}$:

$$\log p(\theta|\mathcal{D}) \approx \log q(\theta) = -\frac{1}{2} (\theta - m)^T H (\theta - m) + \text{constante}$$

Usamos a série de Taylor para aproximar a distribuição *a posteriori* por $q(\theta) = \mathcal{N}(\theta = \mu, \sigma^2 = H^{-1})$, normal multivariada.

Inferência Variacional: aproximar $p(\theta|\mathcal{D})$ por $q(\theta)$. Queremos minimizar uma medida de discrepância entre as duas distribuições, a divergência D_{KL}:

$$D_{\text{KL}}(q||p) = \mathbb{E}_{\theta \sim q} \left[\log \frac{q(\theta)}{p(\theta)} \right] = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta$$

Obtemos a aproximação ótima $\hat{q}$. É zero somente quando $q = p$. $p(\theta)$ na equação acima é a posteriori $p(\theta|\mathcal{D})$. Reescrevemos:

$$\mathbb{E}_{\theta \sim q}[\log q(\theta)] + \mathbb{E}_{\theta \sim q}[\log p(\mathcal{D})] - \mathbb{E}_{\theta \sim q}[\log p(\mathcal{D}|\theta)] - \mathbb{E}_{\theta \sim q}[\log p(\theta)]$$

$\log p(\mathcal{D})$ é constante em relação ao modelo, então minimizar D_{KL} é maximizar $L(q) = \mathbb{E}_{\theta\sim q}[\log p(\mathcal{D}|\theta)] + \mathbb{E}_{\theta\sim q}[\log p(\theta)] - \mathbb{E}_{\theta\sim q}[\log q(\theta)]$.

Escolhemos um espaço de parâmetros Ω que maximiza $L(q)$. Calculamos o gradiente por amostragem, onde utilizamos uma variável aleatória que não depende de Ω (por exemplo, um ruído $\epsilon \sim N[0,1]$) e aplicamos uma transformação $g(\epsilon;(\mu,\sigma^2)) = \epsilon\sigma + \mu$, que depende de Ω. Aproximamos a distribuição *a posteriori* por uma normal multivariada. A fórmula de g varia conforme a distribuição de q (Normal, nesse caso).

SELEÇÃO DE MODELOS

Dilema viés-variância: Queremos um modelo que minimiza a função de perda l média $\mathbb{E}_x[l(h(x), f(x))]$. Suponha função de perda como erro quadrático, então $\mathbb{E}_x[(h(x) - f(x))^2]$.

Queremos um caso geral que não dependa dos dados. Calculamos o valor esperado do erro tratando $\mathcal{D}$ como variável aleatória: $\mathbb{E}_{x,\mathcal{D}}[(h_{\mathcal{D}}(x) - f(x))^2]$.

Abrindo a expressão e somando $\mathbb{E}_{x,\mathcal{D}}[h_{\mathcal{D}}(x)] - \mathbb{E}_{x,\mathcal{D}}[h_{\mathcal{D}}(x)]$, manipulamos até obter

$$\mathbb{E}_x[\underbrace{\mathrm{Var}_{\mathcal{D}}[h_{\mathcal{D}}(x)]}_{\text{Variância}}] + \mathbb{E}_x[(\underbrace{\mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x) - f(x)]}_{\text{Viés}})^2]$$

Se as respostas de teste $f(x)$ fossem corrompidas por um ruído aditivo aleatório, i.e., $f'(x) = f(x) + \epsilon$, obteríamos:

$$\mathbb{E}_x \underbrace{[\mathrm{Var}_{\mathcal{D}}[h_{\mathcal{D}}(x)]]}_{\text{Variância}} + \mathbb{E}_x \underbrace{[(\mathbb{E}_{\mathcal{D}}[f(x) - h_{\mathcal{D}}(x)])^2]}_{\text{Viés}} + \mathbb{E}_{x,\epsilon} \underbrace{[\epsilon^2]}_{\text{Ruído}}$$

Métodos mais flexíveis → maior variância no processo de aprendizado.

Métodos mais simples → baixa variância, mas podem apresentar alto viés.

***Underfitting*:** modelo muito simples para capturar a complexidade dos dados. O modelo não aprende a relação entre as variáveis e generaliza mal.

***Overfitting*:** modelo muito complexo para o problema, se ajusta demais aos dados e os memoriza ao invés de aprender a relação entre as variáveis. Vai bem melhor no treino do que no teste.

Regularização L_2: adicionamos $c||\theta||_2^2$ ao objetivo de aprendizado, i.e., $\arg\min_{\theta} \mathcal{L}_{\mathcal{D}}(\theta) + c||\theta||_2^2$.

RELAÇÕES IMPORTANTES

$D_{\mathrm{KL}}(p||q) = H(p,q) - H(p)$

$H(p) = -\sum_{i=1}^{N} p_i \log p_i$

$\mathrm{cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$

$\mathrm{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

$\mathbb{E}[X] = \mu$, $\mathbb{E}[X^2] = \sigma^2 + \mu^2$

$\sum(y_n - x_n^T\theta)^2 = ||Y - X\theta||^2$

$||u|| = u^T u$

Derivadas de matrizes e vetores

$\frac{\partial x^T a}{\partial x} = \frac{\partial a^T x}{\partial x} = a$; $\frac{\partial a^T X b}{\partial X} = ab^T$

$\frac{\partial a^T X^T b}{\partial X} = ba^T$; $\frac{\partial a^T X a}{\partial X} = \frac{\partial a^T X^T a}{\partial X} = aa^T$

Suponha W uma matriz simétrica:

$\frac{\partial}{\partial s}(x - As)^T W(x - As) = -2A^T W(x - As)$

$\frac{\partial}{\partial x}(x - s)^T W(x - s) = 2W(x - s)$

$\frac{\partial}{\partial s}(x - s)^T W(x - s) = -2W(x - s)$

$\frac{\partial}{\partial x}(x - As)^T W(x - As) = 2W(x - As)$

$\frac{\partial}{\partial A}(x - As)^T W(x - As) = -2W(x - As)s^T$

Gradiente e Hessiana

$f = x^T Ax + b^T x$; $\nabla_x f = \frac{\partial f}{\partial x} = (A + A^T)x + b$; $\frac{\partial^2 f}{\partial x \partial x^T} = A + A^T$

Uma matriz real M de ordem $n \times n$ é definida positiva se $a^T Ma > 0$ para todos os vetores a com entradas reais.

Uma matriz definida positiva pode ser descrita como QAQ^T, onde Q é ortogonal e A diagonal.

A soma de uma matriz semidefinida positiva com uma matriz definida positiva é positiva definida.

A desigualdade de Jensen diz que $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$ para qualquer f convexa. $f(x) = \log x$ é côncava, então usamos $-f(x)$ para aplicar a desigualdade. Em particular, na Divergência KL, $D_{\mathrm{KL}}(q||p)$ se torna $D_{\mathrm{KL}}(p||q)$.

Série de Taylor:

$f(\theta) \approx f(m) + (\theta - m)^T \nabla_\theta f(m) + \frac{1}{2}(\theta - m)^T \nabla_\theta^2 f(m)(\theta - m)$, $f(\theta) = \log p(\theta|\mathcal{D})$.

Propriedades da sigmoide:

$(-t) = 1 - t$; $\frac{\partial \sigma(t)}{\partial t} = \sigma(t)\sigma(-t)$; $\int \sigma(t)dt = \log \sigma(-t) + c$.

Truque da reparametrização:

$X \sim \mathcal{N}(\mu, \Sigma)$.

$X = \mu + L^T Z$, onde $L^T L = \Sigma$.

Mostra-se isso com: $\mathbb{E}[(x - \mu)(x - \mu)^T] = \mathbb{E}[(L^T Z)(L^T Z)^T] = \mathbb{E}[L^T ZZ^T L] = L^T \mathbb{E}[ZZ^T]L = L^T L = \Sigma$.

Manipulação útil: $\prod_{j=1}^{N}\prod_{i=1}^{k} \theta_i^{[y_j=k]} = \prod_{i=1}^{k} \theta_i^{\sum_{j=1}^{N} I(y_j=k)}$

Gaussiana Multivariada:

$p(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left[-\frac{1}{2}(x - m)^T \Sigma^{-1}(x - m)\right]$

$\nabla_x p = -p(x)\sigma^{-1}(x - m)$;

$\nabla_x^2 p = p(x)(\sigma^{-1}(x - m)(x - m)^T \Sigma^{-1} - \Sigma^{-1})$

$l(\theta, \sigma^2|Y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{||\hat{X}\theta - Y||^2}{2\sigma^2}$

$\nabla_\theta l = -\frac{X^T(X\theta - Y)}{\sigma^2}$; $\nabla_\theta^2 l = -\frac{X^T X}{\sigma^2}$;

$\hat{\theta} = (X^T X)^{-1} X^T Y$