

Teorema de Bayes

1) Probabilidade Básica

Temos um experimento aleatório com um espaço de resultados Ω . Um evento A é um subconjunto de Ω . A probabilidade é uma função $P(\cdot)$ que associa um número em $[0, 1]$ a cada evento.

Regras fundamentais:

1) $0 \leq P(A) \leq 1$ para qualquer evento A .

2) $P(\Omega) = 1$.

3) Se A e B são eventos mutuamente exclusivos, então

$$P(A \cup B) = P(A) + P(B)$$

Em geral: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

2) Variável aleatória discrta e contínua

Uma variável aleatória X é uma função que associa cada resultado a um número real.

• discrta: assume valores em um conjunto contável (por exemplo, $0, 1, 2, \dots$). Definimos a função de massa de probabilidade (pmf): $p(x) = P(X = x)$

Propriedades: $p_X(x) \geq 0$ e $\sum_x p_X(x) = 1$

• contínua: assume valores em intervalos reais. Definimos a densidade de probabilidade (pdf): $f_X(x) \geq 0$ com

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

A probabilidade de X cair em um intervalo $[a, b]$ é

$$P(a \leq X \leq b) = \int_a^b f_X(x) dx$$

3) Esperança, variância e desvio padrão

Seja X uma variável real contínua.

• ESPERANÇA

discreta: $E[X] = \sum x p_X(x)$

contínuo: $E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$

• VARIÂNCIA

$$\text{Var}(x) = E[(x - E[x])^2]$$

ou

$$\text{Var}(x) = E[x^2] - (E[x])^2$$

• DESVIO PADRÃO

$$\sigma_x = \sqrt{\text{Var}(x)}$$

4) Probabilidade condicional e independência

A probabilidade condicional de um evento A dado que B ocorreu ($P(A|B) > 0$) é

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Dois eventos A e B são independentes se

$$P(A \cap B) = P(A)P(B)$$

O que é equivalente a dizer

$$P(A|B) = P(A) \quad \text{e} \quad P(B|A) = P(B)$$

5) Probabilidade condicional em termos de densidade (caso contínuo)

Para variáveis aleatórias contínuas X e Y, com densidade conjunta $f_{X,Y}(x,y)$, a densidade condicional de X dado $Y=y$ (assumindo $f_Y(y) > 0$) é

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

A densidade marginal de Y é obtida integrando a conjunta:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx$$

No caso direto, a lógica é a mesma, trocando integrais por somas. Por exemplo:

$$P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

6) Notação $P(x|Y)$, $P(y|x, D)$

- $P(x|Y)$ ou $P(X = x | Y = y)$ significa probabilidade condicional de X assumindo um certo valor, dado um valor específico de Y .
- $P(y|x, D)$
Aqui
 - y é um valor possível da variável de saída
 - x é o valor ou vetor de entradas
 - D é o conjunto de dados ou conhecimento prévio do modelo

"Probabilidade de observar y dado que a entrada é x e dado o conjunto de dados ou informações D ".

7) y vs \hat{y}

- y
 - pode representar uma variável aleatória de saída
 - valor observado real

→ \hat{y}

- representa o valor estimado ou previsto por um modelo

Teorema de Bayes

É uma consequência direta da definição de probabilidade condicional.

Pela definição de probabilidade condicional, temos:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} ; P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Mesmo numerador:

$$1) P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$$2) P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes

A → hipótese

B → dados observados

$$P(w|D) = \frac{P(D|w)P(w)}{P(D)}$$

posterior = likelihood · prior
evidence

1. $P(w | D)$ - posterior

E' a probabilidade da hipótese θ depois de ver os dados
E' o objeto central da inferência Bayesiana

2. $P(D|w)$ - likelihood (verossimilhança)

Mede, quanto os dados observados são "prováveis" se aquela hipótese fosse verdadeira

Não é uma probabilidade sobre w , é uma função dos dados parametrizada por w .

3. $P(w)$ - prior

E' o conhecimento prévio sobre w antes de obter qualquer dado.

Pode expressar crenças, suposições ou distribuições desejadas.

4. $P(D)$ - evidência ou marginal likelihood

E' um termo de normalização que garante que a soma (ou integral) do posterior seja 1.

$$P(D) = \int P(D|w)P(w)dw$$

Esse termo é difícil de calcular na prática, porque é uma integral alta-dimensional sobre todos os pesos da rede, por isso usamos aproximações variacionais.

Weight Uncertainty in Neural Networks [1]

1) A integral impossível do Bayes

No nível mais fundamental, as redes bayesianas querem aprender não um único peso w , mas uma distribuição sobre os pesos. A regra vem do Teorema de Bayes:

$$P(w|D) = \frac{P(D|w)P(w)}{P(D)}$$

$\xrightarrow{\text{likelihood}}$
 $\xrightarrow{\text{prior}}$

↓
posterior

O problema está no denominador:

$$P(D) = \int P(D|w)P(w) dw$$

Essa integral tem de "somar" a probabilidade do dataset inteiro sobre todos os possíveis valores de todos os pesos da rede.

Uma rede neural típica tem milhões ou milhões de pesos. A integral não é unidimensional nem bidimensional — é uma integral em um espaço de dimensão igual ao número de parâmetros do modelo.

2) Aproximar a Posterior

Já que não podemos calcular $P(w|D)$, introduzimos uma família de distribuições simples chamada $q(w|\theta)$, com parâmetros θ (por exemplo, uma Gaussiana com média

μ e desvio σ).

Inférence Variacional: escolher q do jeito certo para que ela fique o mais parecida possível da posterior verdadeira

O critério é formulado assim:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \text{KL}[q(w|\theta) \| P(w|D)]$$



KL Divergence

A KL Divergence (Kullback - Liebler divergence) é uma medida de distância direcionada entre duas distribuições de probabilidade. Funciona como um custo de desemelhança.

Definição geral:

$$\text{KL}(q \| p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

"Quanto custa usar a distribuição q no lugar da distribuição p ?"

A notação $\text{KL}(q \| p)$ lê-se como "KL de q em direção a p " ou "quanto q é diferente de p ".

O primeiro argumento (antes das barras) é sempre o que você está tentando aproximar ou usar como substituto. O se-

quando é o alvo.

$q \parallel p \rightarrow$ "quão ruim é usar q no lugar de p ?"

A ordem importa: KL é assimétrico:

$$KL(q \parallel p) \neq KL(p \parallel q)$$

No artigo [5] a equação que aparece é

$$KL[q(w|\theta) \parallel P(w|D)]$$

que significa "o quão diferente é a distribuição aproximada dos pesos ($q(w|\theta)$) da verdadeira posterior Bayesiana ($P(w|D)$)."

$q(w|\theta)$: é a distribuição que nós escolhemos, normalmente uma Gaussiana diagonal. Depende de $\theta = (\mu, \sigma)$.

$P(w|D)$: é a distribuição verdadeira dos pesos depois de observar os dados. É o que gostaríamos de saber, mas não sabemos calcular porque exige a integral impossível.

$k(q \parallel P)$: mede o quanto nossa aproximação está errada em relação à posterior verdadeira.

* escolher os parâmetros θ de q para deixar q o mais parecido possível com a posterior verdadeira.

$$\rightarrow \text{minimizar } KL(q(w|\theta) || P(w|D))$$

Se KL for zero, significa que $q(w|\theta) = P(w|D)$, e a aproximação é perfeita.

3) O KL vale a pena: maxima ELBO

Expandido KL , o cálculo mostra que:

$$KL[q(w|\theta) || P(w|D)] = KL[q(w|\theta) || P(w)] - E_{q(w|\theta)}[\log P(D|w)] + \log P(D)$$

O termo $P(D)$ é constante (não depende de θ)

Então basta minimizar:

$$F(D, \theta) = KL[q(w|\theta) || P(w)] - E_{q(w|\theta)}[\log P(D|w)]$$

↳ variational free energy / negative ELBO

4) A esperança difícil

Agora temos a expressão:

$$E_{q(w|\theta)}[\log P(D|w)]$$

Ela depende de w , mas w é amostrado da distribuição variacional, que por sua vez depende de θ . Isso significa que o gradiente que queremos calcular é:

$$\frac{\partial}{\partial \theta} \text{Eq}_{(w|\theta)} [\log P(D|w)]$$

PROBLEMA: não pode derivar diretamente através de uma variável aleatória. Se tentar derivar a densidade q , os gradientes ficam de variação altíssima.

SOLUÇÃO: Reparametrization Trick

5) Reparametrization Trick

1) geramos uma amostra de ruído puro, onde pendente dos parâmetros $\in \sim N(0, I)$

2) transformamos essa amostra em um pés

$$w = \mu + \sigma \cdot \epsilon$$

- w não é mais diretamente sortido
- w é uma função determinística de θ com ruído fixo separado.

Caso Gaussiano $w = \mu + \log(1 + e^P) \circ \epsilon$

onde ρ parametriza o desvio padrão e garante que ele seja positivo.

6) Por que isso resolve tudo?

Por que agora

$$E_{q(w|\theta)}[f(w)] = E_{p(\epsilon)}[f(t(\theta, \epsilon))].$$

A esperança passou a ser sobre um mundo que não depende de θ .

Logo, podemos impulsionar o gradiente para dentro:

$$\frac{\partial}{\partial \theta} E_{\epsilon}[f(t(\theta, \epsilon))] = E_{\epsilon} \left[\frac{\partial f}{\partial w} \frac{\partial w}{\partial \theta} \right]$$

- $t(\theta, \epsilon)$ é uma função diferenciável
- ϵ é constante no sentido do gradiente

7) Ciclo completo

① Bayes exige integrar sobre todos os pesos
→ integral impossível em dimensão altíssima

② Aprimora-se a posterior com $q(w|\theta)$
→ introduz KL e ELBO

③ O ELBO exige calcular uma esperança
→ mas a esperança envolve o corteado de $q(w|\theta)$, impossível de derivar diretamente

④ Entra o Reparametrization Trick
→ transforma o cortejo em função determinística de w + parâmetros

⑤ Agora a esperança é uma esperança sobre w só que passa
→ o gradiente passa por w tranquilamente

⑥ Inféria variacional via backpropagation padrão
→ método escalável

ELBO é Evidence Lower Bound, uma função que substitui o cálculo impossível da posterior Bayesiana, servindo como um valo de otimização. É a função que, quando maximizada, faz a distribuição variacional $q(w|\theta)$ ficar o mais parecida possível da posterior verdadeira $P(w|D)$, sem precisar calcular a integral de Bayes.

REFERÊNCIAS

- [1] C. Blundell, S. Corneilie, K. Kavukcuoglu and D. Wierstra
"Weight Uncertainty in Neural Networks". Proceedings of the
32nd International Conference on Machine Learning, Lille,
France, 2015.

Principal