Exercícios Deep Learning
Aula 15

October 17, 2019

1 Variational autoencoder

1- Explique o que são: Sparse autoencoders, Denoising autoencoders, Contractive autoencoders e Variational autoencoders.

2- O que é um modelo generativo? Por que o variational autoencoder, ao contrário do autoencoder, pode ser considerado um modelo generativo?

3- Especifique a função de perda que é minimizada no variational autoencoder.

4- Dado duas distribuições Bernoulli, P e Q, especifique a divergência de Kullback–Leibler, $D_{\mathrm{KL}}(P||Q)$, entre elas.

5- Dados duas distribuições para $x = (0, 1, 2)$, $P(x) = (0.36, 0.48, 0.16)$ e $Q(x) = (0.333, 0.333, 0.333)$. Calcule a divergência de Kullback–Leibler entre elas. Verifique que a divergência KL não é simétrica, ou seja, $D_{\mathrm{KL}}(P||Q)$ é diferente de $D_{\mathrm{KL}}(Q||P)$.

6- Como é possível fazer o backpropagation através do sampling feito na camada de coding?

7- Considere um variational autoencoder que tem como entrada uma única feature x. Seja $u_\phi(x) = ReLU(2\ ReLU(3x - 1)), \Sigma_\phi(x) = ReLU(ReLU(3x + 1) + 1)$ e $u_\theta(z) = ReLU(-z + 12), \Sigma_\theta(z) = ReLU(2z - 10)$

a) Gere 3 amostras de z a partir de $x = 2$ e $\varepsilon = 0.5, 0, -0.5$.

b) Gere 3 amostras de x a partir das amostras de z geradas no exercício acima (1 para cada z).

8- Na derivação da função ELBO, por que o terceiro fator $KL(q(z|x)||p(z||x))$ é descartado?

9- Utilizando VAEs é possível de maneira simples criar novas amostras apenas modificando cada posição do vetor latente do modelo. Entretanto, ainda que essas posições capturem fatores da imagem (posição, tamanho, rotação), é difícil de se controlar novas amostras geradas, já que os fatores podem ser codificados em múltiplos componentes interdependentes de z. Qual seria uma possível solução para tornar a codificação latente mais controlável, ou seja, gerar codificações em z onde os múltiplos componentes são independentes entre si. Qual é uma possível desvantagem dessa abordagem?

Solução

1- Sparse autoencoder: É adicionado regularização L1 na função de perda para minimizar a dimensão do espaço latente.

Denoising autoencoder: O autoencoder é treinado com entradas adicionadas de um ruído, porém a função de perda considera a distância do valor reconstruído ($\hat{x}$) com a entrada original sem ruído. Dessa forma, o autoencoder aprende a retirar o ruído das entradas.

Contractive autoencoder: A função de perda contém a norma do vetor gradiente para que uma mudança pequena em x produza uma mudança também pequena no espaço latente.

Variational autoencoder: Um modelo generativo onde o espaço latente é na verdade uma distribuição de probabilidades.

2- Um modelo generativo é um modelo que especifica uma distribuição de probabilidades na entrada, dessa forma, é possível gerar novos exemplos de entradas. O variational autoencoder é um modelo generativo pois ele estima uma função $p(x|z)$ que pode ser usada para amostrar novos valores de x.

3- A função de perda do variational autoencoder é a log-verossimilhança negativa com um regularizador. Como não há representações globais compartilhadas por todos os pontos de dados, podemos decompor a função de perda em apenas termos que dependem de um único ponto de dado l_i. A perda total é $\sum_{i=1}^{N} l_i$ para N total de pontos de dados. A função de perda l_i para cada dado de entrado x_i é:

$$l_i(\theta, \phi) = -\mathbb{E}_{z \sim q_\theta(z|x_i)} \left[\log p_\phi \left(x_i|z\right)\right] + \mathbb{KL}\left(q_\theta \left(z|x_i\right) \| p(z)\right) \tag{1}$$

O primeiro termo é a perda de reconstrução do i-ésimo dado de entrada. A esperança é tomada em relação à distribuição do codificador sobre as representações. Este termo incentiva o decodificador a aprender a reconstruir os dados. Se a saída do decodificador não reconstruir bem os dados, estatisticamente dizemos que o decodificador parametriza uma distribuição de probabilidade que não coloca muita probabilidade de massa nos dados verdadeiros.

O segundo termo é a divergência de Kullback-Leibler entre a distribuição do codificador $q_\theta(z|x)$ e $p(z)$. Essa divergência mede quanta informação é perdida ao usar q para representar p. É uma medida de quão próximo q é para p.

4- $D_{\mathrm{KL}}(P||Q) = D_{\mathrm{KL}}(Bernoulli(P)||Bernoulli(Q)) = P\ log(\frac{P}{Q}) + (1 - P)log(\frac{1-P}{1-Q})$

5-

$$
\begin{aligned}
D_{\mathrm{KL}}(P\|Q) &= \sum_{x\in\mathcal{X}} P(x)\ln\left(\frac{P(x)}{Q(x)}\right) \\
&= 0.36\ln\left(\frac{0.36}{0.333}\right) + 0.48\ln\left(\frac{0.48}{0.333}\right) + 0.16\ln\left(\frac{0.16}{0.333}\right) \\
&= 0.0852996
\end{aligned}
$$

$$
\begin{aligned}
D_{\mathrm{KL}}(Q\|P) &= \sum_{x\in\mathcal{X}} Q(x)\ln\left(\frac{Q(x)}{P(x)}\right) \\
&= 0.33\ln\left(\frac{0.333}{0.36}\right) + 0.33\ln\left(\frac{0.333}{0.48}\right) + 0.333\ln\left(\frac{0.333}{0.16}\right) \\
&= 0.097455
\end{aligned}
$$

6- É utilizado um truque de reparametrização: Em vez de amostrarmos $z \sim \mathcal{N}(u_\phi(x), \Sigma_\phi(x))$, amostramos $\varepsilon \sim \mathcal{N}(0,1)$ e fazemos $z = u_\phi(x) + \Sigma_\phi(x)^{1/2}\varepsilon$. Dessa forma o sampling só é feito em ε e o erro não precisa ser propagado nesta parte.

7-
a) $u_\phi(x) = 10, \Sigma_\phi(x) = 8,\ z = \{10+4, 10+0, 10-4\} = \{14, 10, 6\}$
b) $u_\theta(z) = \{0, 2, 6\}, \Sigma_\theta(z) = \{18, 10, 2\}$. Gerar uma amostra de $\mathcal{N}(0,18), \mathcal{N}(2,10)$ e $\mathcal{N}(6,2)$.

8- $p(z|x) = \frac{p(x|z)p(z)}{p(x)}$, como $p(x)$ em geral não pode ser resolvido, $p(z|x)$ também é intratável e não conseguimos calcular a KL envolvendo $p(z|x)$.

9- No VAE tradicional, uma Gaussiana ($p(z) \approx N(0, I)$) é tipicamente usado como a distribuição a priori anterior para z. Note que sob esta distribuição os componentes de z são independentes, que é exatamente a propriedade que gostaríamos que nossa distribuição posteriori aproximada, por exemplo $q(z|x)$, tivesse. Para garantir a independência, multiplicamos o termo de divergência de KL no ELBO por um fator β. A desvantagem é que as imagens ficam consideravelmente mais desfocadas, já que o aumento da divergência de KL reduz a flexibilidade da distribuição a posteriori, ao enfatizar fator de divergência da função de perda.