←

# Variational autoencoders

测验, 8 个问题

✖ **准备好后再次尝试。**

通过所需分数：80% 或更高

每隔 8 小时，您最多可以重新进行 3 次 此测验。

返回到第 5 周

重新测试

✓ 1 / 1 分

1。

Which of the following estimates are unbiased?

☑ ~~$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i,\;\; X_i \overset{i.i.d.}{\sim} p$ for $\mathbb{E}X$.~~

**正确**

☑ ~~$f\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right),\;\; X_i \overset{i.i.d.}{\sim} p$ for $\mathbb{E}f(X)$, where $f$ is a linear function.~~

**正确**

For a linear function, ~~$f\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$ which is an unbiased estimate for $\mathbb{E}f(X)$. (See Lec 1 starting~~ at 2:03.)

□ ~~$f\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right),\;\; X_i \overset{i.i.d.}{\sim} p$ for $\mathbb{E}f(X)$, where $f$ is not linear.~~

**未选择的是正确的**

0.67 / 1 分

2。

In which of the following scenarios probabilistic model has a tractable density function $p(x)$, i.e. it is computationally easy to compute the density at any point $x$?

☑ $x$ is an observable variable in an arbitrary latent variable model $p(x, z) = p(x \mid z)p(z)$

**这个选项的答案不正确**

Computing $p(x)$ involves integrating (or summing if $z$ is descrete) w.r.t. $z$ $p(x) = \int p(x \mid z)p(z)dz$ which can be very complicated.

☑ Distribution is defined according to the chain rule $p(X) = \prod_{i=1}^{D} p_i(x_i \mid x_1, \dots, x_{i-1})$, $X = (x_1, \dots, x_D)$ with tractable conditional distributions.

**正确**

Indeed in this case we can always compute the density using the chain rule.

□ $x$ is defined by a smooth transformation $f : \mathbb{R}^d \to \mathbb{R}^d$ of a random vector $z \sim \mathcal{N}(0, I)$: $x = f(\varepsilon)$.

**未选择的是正确的**

✗ 0 / 1 分

3。

Consider two different choices for the family of variational distributions for training a VAE:

1) $q(z_i \mid x_i, w) = \mathcal{N}(z_i \mid \mu(x_i, w), \mathrm{diag}(\sigma^2(x_i, w)))$, where $\mu(\cdot, w)$ and $\sigma(\cdot, w)$ are deep neural networks with parameters $w$.

2) For each $x_i$ the approximate posterior distribution over latent variable $z_i$ is defined individually as a Gaussian distribution $q_i(z_i \mid x_i, \mu_i, \sigma_i) = \mathcal{N}(z_i \mid \mu_i, \mathrm{diag}(\sigma_i^2))$.

In which case, you would expect to get higher (better) variational lower bound on the training set? What about the test set?

○ 2 is better for both training and test sets.

○ 2 is better on the training set, 1 is better on the test set.

● 1 is better for both training and test sets.

**这个选项的答案不正确**
Note that option 2 is more flexible, so at least on the training set should achieve better loss.

○ 1 is better on the training set, 2 is better on the test set.

---

✓ 1 / 1 分

4。
Suppose the class of the approximate posterior distributions is flexible enough to capture any distribution. The evidence lower bound is known to achieve the optimal value with respect to the variational distribution when the variational distribution coincides with the true posterior distribution $q(z \mid x, w) = p(z \mid x)$.

Imagine that you train VAE by optimizing the following loss: $\sum_i \mathbb{E}_{q(z|x_i,w)} \log[p(x_i \mid z)p(z)]$, which is the usual variational lower bound $\sum_i \mathbb{E}_{q(z|x_i,w)} \log p(x_i \mid z) - \mathcal{KL}(q(z \mid x_i, w) \,\|\, p(z))$ without the entropy term $E_{q(z|x_i,w)} \log q(z \mid x_i, w)$. Which variational distribution you will obtain after training?

○ None of the above.

○ True posterior distribution $p(z \mid x)$.

○ A delta function concentrated in the mode of the posterior $\mathrm{argmax}_z p(z \mid x)$.

● A delta function concentrated in the mode of the joint distribution $\mathrm{argmax}_z p(z, x)$.

**正确**
Indeed, without the entropy term, nothing will stop the variational distribution from collapsing to a delta function that maximizes the objective. If there are multiple modes, the variational distribution can become a mixture of them.

---

✗ 0 / 1 分

5。
Suppose that a random variable $z_j$ does not contribute to the value of decoder. That is, no matter what the value $z_j$ takes, the output of decoder is the same. What will be the distribution $q(z_j \mid x, w)$ after training?

○ The component will be distributed according to the prior distribution $\mathcal{N}(z_j \mid 0, 1)$

○ None of the above.

● Distribution $q(z_j \mid x, w)$ will not change during the training.

**这个选项的答案不正确**
Although the reconstruction term of the loss doesn't depend on the distribution, the KL term does and will force the distribution to change.

---

0.50 / 1
分

6。

Recall the log derivative trick (or as it is sometimes called, REINFORCE algorithm). It estimates the gradient $\nabla_w \mathbb{E}_{q(z|w)} f(z)$ by using the following derivation $\nabla_w \mathbb{E}_{q(z|w)} f(z) = \int \nabla_w q(z \mid w) f(z) dz = \int \frac{q(z|w)}{q(z|w)} \nabla_w q(z \mid w) f(z) dz =$

$$\mathbb{E}_{q(z|w)} \nabla_w \log q(z \mid w) f(z)$$

and then builds an unbiased estimate $\mathbb{E}_{q(z|w)} \nabla_w \log q(z \mid w) f(z) \approx \nabla_w \log q(\hat{z} \mid w) f(\hat{z})$, where $\hat{z} \sim q(z \mid w)$.

Which conditions are necessary for the log derivative trick to work?

□ One has to represent the variational approximation $q(z \mid w)$ with a function $g$ and a random variable $\varepsilon$ such that $z = g(\varepsilon, w)$.

**未选择的是正确的**

☑ Density function $q(z \mid w)$ must be differentiable with respect to the random variable value $z$.

**这个选项的答案不正确**

The log-derivative trick doesn't use the derivative w.r.t. $z$, so it is applicable for example to Laplace distribution, which has a non-differentiable peak.

□ Density function $q(z \mid w)$ must be differentiable with respect to the distribution parameters $w$.

**这应该被选择**

□ Function $f(z)$ should be differentiable w.r.t. $z$.

**未选择的是正确的**

0.50 / 1
分

7。
Recall the derivation of the reparametrization trick. It estimates the gradient $\nabla_w \mathbb{E}_{q(z|w)} f(z)$ by introducing a function $g$ such that $g(\varepsilon, w)$ is distributed according to $q(z \mid w)$ where the distribution of $\varepsilon$ doesn't depend on parameters $w$. Then, $\nabla_w \mathbb{E}_{q(z|w)} f(z) = \nabla_w \mathbb{E}_{p(\varepsilon)} f(g(\varepsilon, w)) = \mathbb{E}_{p(\varepsilon)} \nabla_w f(g(\varepsilon, w))$ which can be easily estimated in an unbiased way as $\nabla_w f(g(\hat{\varepsilon}, w))$ where $\hat{\varepsilon} \sim p(\varepsilon)$.

Which conditions are necessary for the reparametrization trick to work?

☑ One has to represent the variational approximation $q(z \mid w)$ with a function $g$ and a random variable $\varepsilon$ such that $z = g(\varepsilon, w)$.

**正确**

This is the first step of the reparameterization trick.

□ Density function $q(z \mid w)$ must be differentiable with respect to the random variable value $z$.

**未选择的是正确的**

□ $g(\varepsilon, w)$ must be differentiable with respect to $w$.

**这应该被选择**

□ Function $f(z)$ should be differentiable w.r.t. $z$.

**这应该被选择**

✓ 1 / 1 分

8。

Which of the following distribution families can be used in the reparametrization trick?

☐ Bernoulli distribution.

**未选择的是正确的**

☑ Multivariate normal distribution with full convariance matrix.

**正确**

$z = g(\varepsilon, \mu, \Sigma) = \Sigma^{0.5}\varepsilon + \mu$, where $\varepsilon$ follows standard normal distribution and $\Sigma^{0.5}$ is the square root of the matrix $\Sigma$, i.e. a matrix $B$ such that $BB = \Sigma$. Latent variable $z$ expressed this way follows $z \sim \mathcal{N}(\mu, \Sigma)$.

Note that $g(\varepsilon, \mu, \Sigma)$ is differentiable w.r.t. both the parameter-vector $\mu$ and the parameter-matrix $\Sigma$.

☑ Multivariate normal distribution with diagonal covariance matrix.

**正确**

$z = g(\varepsilon, \mu, \sigma) = \varepsilon \odot \sigma + \mu$, where $\varepsilon$ follows standard normal distribution and $\odot$ is element-wise multiplication. Latent variable $z$ expressed this way follows $z \sim \mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$.

Note that $g(\varepsilon, \mu, \sigma)$ is differentiable w.r.t. both parameter-vectors $\mu$ and $\sigma$.