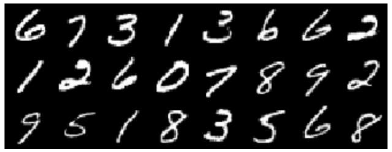
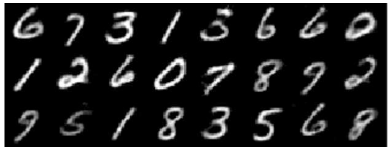
Introduction to the problem setup

What do we mean by modeling the unknown distribution $P(X)$?





Introduction to the problem setup

Problem setup:

- Input: Data set $D = \{x^i\}_{i=1}^N$, where $x^i \in \mathcal{X}$ denotes the i-th data sample or data point.
- $\mathcal{X}$ is an appropriate input space.
- Examples of $\mathcal{X}$:
  - Set of images (*e.g.* digits, faces, animals, etc.) or videos.
  - Set of vector-valued data or matrix-valued data or tensor-valued data.
  - Set of natural language sentences.
  - Set of documents (*e.g* newspaper articles, books).
  - Set of software programs.

Introduction to the problem setup

What do we mean by modeling the unknown distribution $P(X)$?

- For realizations x of X which are similar to the data samples in $D = \{x^i\}_{i=1}^N$, we want $P_X(x)$ to take higher values (*e.g.* $P_X(x) \approx 1$).
- For realizations x of X which are **not** similar to the data samples in $D = \{x^i\}_{i=1}^N$, we want $P_X(x)$ to take smaller values (*e.g.* $P_X(x) \approx 0$).

Introduction to the problem setup

Problem setup:

- Input: Data set $D = \{x^i\}_{i=1}^N$, where $x^i \in \mathcal{X}$ denotes the i-th data sample or data point.
- Assumption: The data samples are generated from some unknown probability distribution denoted by $P(X)$ (also denoted by $P_X(x)$).
  - Note: X is a random variable and x is a realization of X.
- Aim: To model the unknown distribution $P(X)$ using the observed data samples $D = \{x^i\}_{i=1}^N$.

Introduction to the problem setup

Several ways to model $P(X)$:

- Density estimation techniques
  - Kernel density estimation
  - Spectral density estimation
  - Non-parametric density estimation
- Histogram fitting
- Generative models

Generative models

Recall:

- Input: Data set $D = \{x^i\}_{i=1}^N$, where $x^i \in \mathcal{X}$ denotes the i-th data sample or data point.
- Assumption: The data samples are generated from some unknown probability distribution denoted by $P_X(x)$.

Generative models

Recall:

- Input: Data set $D = \{x^i\}_{i=1}^N$, where $x^i \in \mathcal{X}$ denotes the i-th data sample or data point.
- Assumption: The data samples are generated from some unknown probability distribution denoted by $P_X(x)$.

Generative Model:

- A machine learning model of the unknown $P_X(x)$, given the data $D = \{x^i\}_{i=1}^N$, where $x^i \in \mathcal{X}$.

Generative models

- Input: Data set $D = \{x^i\}_{i=1}^N$, where $x^i \in \mathcal{X}$ denotes the i-th data sample or data point.
- Modeling $P_X(x)$ is usually a complicated task.
  - Suppose that $x \in \mathbb{R}^d$ or $x \in \mathbb{R}^{m \times n}$ (that is, x is vector-valued or matrix-valued).

Generative models

- Input: Data set $D = \{x^i\}_{i=1}^N$, where $x^i \in \mathcal{X}$ denotes the i-th data sample or data point.
- Modeling $P_X(x)$ is usually a complicated task.
  - Suppose that $x \in \mathbb{R}^d$ or $x \in \mathbb{R}^{m\times n}$ (that is, x is vector-valued or matrix-valued).
  - Correlations between various components of x might be difficult to model.
- To model $P_X(x)$ we use a trick:
  - Marginal distribution:

$$\begin{aligned} P_X(x) &= \int P_{(X,Z)}(x,z)dz \\ &= \int P_{X|Z}(x|z)P_Z(z)dz \text{ (By definition of conditional probability)} \end{aligned}$$

Generative models

- Marginal distribution:

$$\begin{aligned} P_X(x) &= \int P_{(X,Z)}(x,z)dz \\ &= \int P_{X|Z}(x|z)P_Z(z)dz \end{aligned}$$

- **Why do we need this marginal distribution?**
  - We introduce a new random variable Z.
  - Z is called a latent variable.
  - Z is usually under our control.
  - By suitable choice of Z with a known **prior** distribution $P_Z(z)$ we can try to model $P_X(x)$ if we can effectively model the **conditional** distribution $P_{X|Z}(x|z)$.

Generative models

- Marginal distribution:

$$\begin{aligned} P_X(x) &= \int P_{(X,Z)}(x,z)dz \\ &= \int P_{X|Z}(x|z)P_Z(z)dz \end{aligned}$$

Requirements:

- A suitable choice for **prior** distribution $P_Z(z)$.
- How to model the **conditional** distribution $P_{X|Z}(x|z)$?

Catch:

- Computing the integral is generally **intractable**.
- Need to use computationally intensive Markov-Chain Monte Carlo techniques to estimate the integral.

Generative models

A different approach:

- True posterior:

$$P_{Z|X}(z|x) = \frac{P_{X|Z}(x|z)P_Z(z)}{P_X(x)} \text{ (By Bayes' Theorem \& law of total proabability)}$$

can be used to model $P_X(x)$.

Recap of Variational Bayes

- True posterior:

$$P_{Z|X}(z|x) = \frac{P_{X|Z}(x|z)P_Z(z)}{P_X(x)}$$

  can be used to model $P_X(x)$.
- Computing $P_{Z|X}(z|x)$ is intractable in general.
- Use a customized distribution $Q_{Z|X}(z|x)$ (called recognition model) to approximate $P_{Z|X}(z|x)$.

Recap of Variational Bayes

- Error of approximation between $P_{Z|X}$ and $Q_{Z|X}$ can be computed using the Kullback-Leibler (or KL)-Divergence:

$$\begin{aligned}
&KL(Q_{Z|X}(z|x)||P_{Z|X}(z|x)) \\
&= \int \log \frac{Q_{Z|X}(z|x)}{P_{Z|X}(z|x)} Q_{Z|X}(z|x)dz \\
&= E_{Z\sim Q}\left[\log Q_{Z|X}(z|x) - \log P_{Z|X}(z|x)\right] \\
&= E_{Z\sim Q}\left[\log Q_{Z|X}(z|x) - \log \frac{P_{X|Z}(x|z)P_Z(z)}{P_X(x)}\right] \\
&= E_{Z\sim Q}\left[\log Q_{Z|X}(z|x) - \log P_{X|Z}(x|z) - \log P_Z(z)\right] + \log P_X(x)
\end{aligned}$$

Recap of Variational Bayes

- Thus we have:

$$\begin{aligned}
&KL(Q_{Z|X}(z|x)||P_{Z|X}(z|x)) \\
&= E_{Z\sim Q}\left[\log Q_{Z|X}(z|x) - \log P_{X|Z}(x|z) - \log P_Z(z)\right] + \log P_X(x)
\end{aligned}$$

- Rearranging, we get:

$$\begin{aligned}
&\log P_X(x) - KL(Q_{Z|X}(z|x)||P_{Z|X}(z|x)) \\
&= E_{Z\sim Q}\left[\log P_Z(z) - \log Q_{Z|X}(z|x) + \log P_{X|Z}(x|z)\right] \\
&= E_{Z\sim Q}\left[-\log \frac{Q_{Z|X}(z|x)}{P_Z(z)} + \log P_{X|Z}(x|z)\right] \\
&= E_{Z\sim Q}\left[\log P_{X|Z}(x|z)\right] - E_{Z\sim Q}\left[\log \frac{Q_{Z|X}(z|x)}{P_Z(z)}\right] \\
&= E_{Z\sim Q}\left[\log P_{X|Z}(x|z)\right] - KL(Q_{Z|X}(z|x)||P_Z(z))
\end{aligned}$$

Recap of Variational Bayes

Recall our idea:

- Use a customized distribution $Q_{Z|X}(z|x)$ (called recognition model) to approximate $P_{Z|X}(z|x)$.

Objective:

$$\log P_X(x) - KL(Q_{Z|X}(z|x)||P_{Z|X}(z|x)) = E_{Z\sim Q}\left[\log P_{X|Z}(x|z)\right] - KL(Q_{Z|X}(z|x)||P_Z(z))$$

Aim: To maximize the objective.

Note:

- $\log P_X(x)$ denotes the log likelihood, which we wanted to model.
- $KL(Q_{Z|X}(z|x)||P_{Z|X}(z|x))$ denotes the dissimilarity between the recognition distribution Q and the true posterior $P_{Z|X}(z|x)$.
- The KL term acts like a regularizer.

Recap of Variational Bayes

Recall: Our aim is to maximize objective:

$$\log P_X(x) - KL(Q_{Z|X}(z|x)||P_{Z|X}(z|x)) = E_{Z\sim Q}\left[\log P_{X|Z}(x|z)\right] - KL(Q_{Z|X}(z|x)||P_Z(z))$$

In the presence of a dataset D, our objective would become:

$$\begin{aligned} \max \ & E_{X\sim D}\left[\log P_X(x) - KL(Q_{Z|X}(z|x)||P_{Z|X}(z|x))\right] \\ &= E_{X\sim D}\left[E_{Z\sim Q}\left[\log P_{X|Z}(x|z)\right] - KL(Q_{Z|X}(z|x)||P_Z(z))\right] \end{aligned}$$

For a sample x^i from D, the corresponding objective term is:

$$\mathcal{L}(\theta,\phi;x^i) = E_{Z\sim Q}\left[\log P_{X|Z}(x^i|z)\right] - KL(Q_{Z|X}(z|x^i)||P_Z(z))$$

- For data set $D = \{x^i\}_{i=1}^N$ and a randomly chosen minibatch $\mathcal{B}$ of size M, we can find $\mathcal{L}(\theta,\phi;\mathcal{B}) = \frac{M}{N}\sum_{x\in\mathcal{B}}\mathcal{L}(\theta,\phi;x)$.

Recap of Variational Bayes

To maximize Objective:

$$\log P_X(x) - KL(Q_{Z|X}(z|x)||P_{Z|X}(z|x)) = E_{Z\sim Q}\left[\log P_{X|Z}(x|z)\right] - KL(Q_{Z|X}(z|x)||P_Z(z))$$

- We parametrize P using θ.
- We parametrize Q using ϕ.

Thus we get:

$$\begin{aligned} \log P_X(x;\theta) - KL(Q_{Z|X}(z|x;\phi)||P_{Z|X}(z|x;\theta)) = {} & E_{Z\sim Q}\left[\log P_{X|Z}(x|z;\theta)\right] \\ & - KL(Q_{Z|X}(z|x;\phi)||P_Z(z;\theta)) \end{aligned}$$

VAE

- Major assumption: $P_Z(z) \approx \mathcal{N}(0, I)$.
- We assume $Q_{Z|X}(z|x^i;\phi) = \mathcal{N}(z;\mu^i,(\sigma^i)^2 I)$, where $\mathcal{N}(z;\mu^i,(\sigma^i)^2 I)$ denotes the normal distribution with mean $\mu^i = \mu(x^i)$ and covariance matrix $(\sigma^i)^2 I$ with $(\sigma^i)^2 = \sigma^2(x^i)$.
- We can adopt reparametrization trick using $G(x^i,\epsilon^l;\phi) = \mu^i + \sigma^i \odot \epsilon^l$ where $\epsilon^l \sim \mathcal{N}(0, I)$ and $\odot$ denotes the elementwise multiplication.
- Then we can write:

$$\mathcal{L}(\theta,\phi;x^i) \approx \frac{1}{2}\sum_{j=1}^{d}\left(1 + \log((\sigma_j^i)^2) - (\mu_j^i)^2 - (\sigma_j^i)^2\right) + E_{Z\sim Q}\left[\log P_{X|Z}(x^i|z)\right]$$

Recap of Variational Bayes

Coding theory perspective:

- $Q_{Z|X}(z|x;\phi)$ is called a probabilistic encoder since given a sample x, Q encodes it into a distribution.
- $P_{X|Z}(x|z;\theta)$ is called a probabilistic decoder since given a latent variable z, P produces a distribution over corresponding values of x.
- Hence the methodology is called auto-encoding variational Bayes (AEVB).

Generative models - Variational Bayes Approach

Requirement:

- Given an approximate posterior $Q_{Z|X}(z|x)$, we need to sample $Z \sim Q$.

Caveat:

- The approximate posterior $Q_{Z|X}(z|x)$ might not be differentiable.

Reparametrization trick for sampling Z

- Assume a differentiable $G(\epsilon, x; \phi)$ and sample $Z \sim G$.
- **Note:** We have now introduced a new variable ϵ.
- **Assumption:** $\epsilon \sim p(\epsilon)$.

VAE

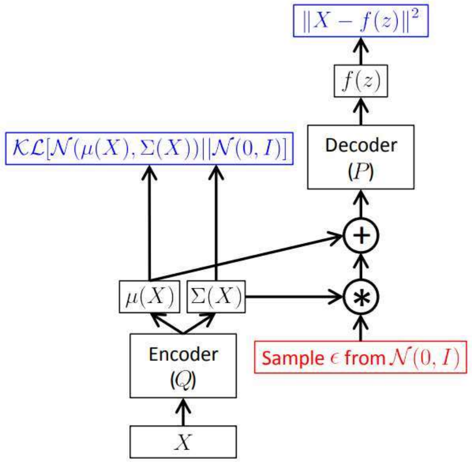
- Hence the overall loss becomes:

$$\mathcal{L}(\theta, \phi; x^i) \approx \frac{1}{2} \sum_{i=1}^{d} \left(1 + \log((\sigma_j^i)^2) - (\mu_j^i)^2 - (\sigma_j^i)^2\right) + \frac{1}{L} \sum_{l=1}^{L} \left(\log P_{X|Z}(x^i|z^{i,l})\right)$$

where $z^{i,l} \sim G(\epsilon^{i,l}, x^i; \phi)$ and $\epsilon^{i,l} \sim p(\epsilon)$.

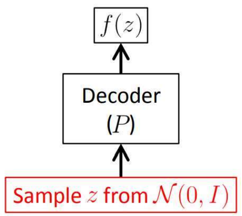
VAE

Training a VAE:



VAE

Testing VAE:



- Remove the encoder
- Sample $z \sim \mathcal{N}(0, I)$.
- Generate sample using decoder.