Bayesian Deep Neural Networks

Summary

Sungjoon Choi

February 9, 2018

Seoul National University

Table of contents

Measure theory

Measure theory

- **set function**: a function assigning a number of a set (example: cardinality, length, area).
- σ-**field** $\mathcal{B}$: a collection of subsets of U such that (axioms)
  1. $\varnothing \in \mathcal{B}$ (empty set is included.)
  2. $B \in \mathcal{B} \Rightarrow B^c \in \mathcal{B}$ (closed under set complement.)
  3. $B_i \in \mathcal{B} \Rightarrow \cup_{i=1}^{\infty} B_i \in \mathcal{B}$ (closed under countable union.)

Measure theory

- properties of σ-**field** $\mathcal{B}$
  1. $U \in \mathcal{B}$ (entire set is included.)
  2. $B_i \in \mathcal{B} \Rightarrow \cap_{i=1}^{\infty} B_i \in \mathcal{B}$ (closed under countable intersection)
  3. 2^U is a σ-field.
  4. $\mathcal{B}$ is either finite or uncountable, never denumerable.
  5. $\mathcal{B}$ and $\mathcal{C}$ are σ-fields $\Rightarrow \mathcal{B} \cap \mathcal{C}$ is a σ-field but $\mathcal{B} \cup \mathcal{C}$ is not.
     - $\mathcal{B} = \{\varnothing, \{a\}, \{b, c\}, \{a, b, c\}\}$
     - $\mathcal{C} = \{\varnothing, \{a, b\}, \{c\}, \{a, b, c\}\}$
     - $\mathcal{B} \cap \mathcal{C} = \{\varnothing, \{a, b, c\}\}$
       (this is a σ-field)
     - $\mathcal{B} \cup \mathcal{C} = \{\varnothing, \{a\}, \{c\}, \{a, b\}, \{b, c\}, \{a, b, c\}\}$
       (this is not a σ-field as $\{a, c\} = \{a\} \cap \{c\}$ is not included.)
- $\sigma(\mathcal{C})$ is called the σ-field **generated** by $\mathcal{C}$.

Probability

Probability

- The **random experiment** should be well defined.
- The **outcomes** are all the possible results of the random experiment each of which canot be further divided.
- The **sample point** w: a point representing an outcome.
- The **sample space** Ω: the set of all the sample points.

Probability

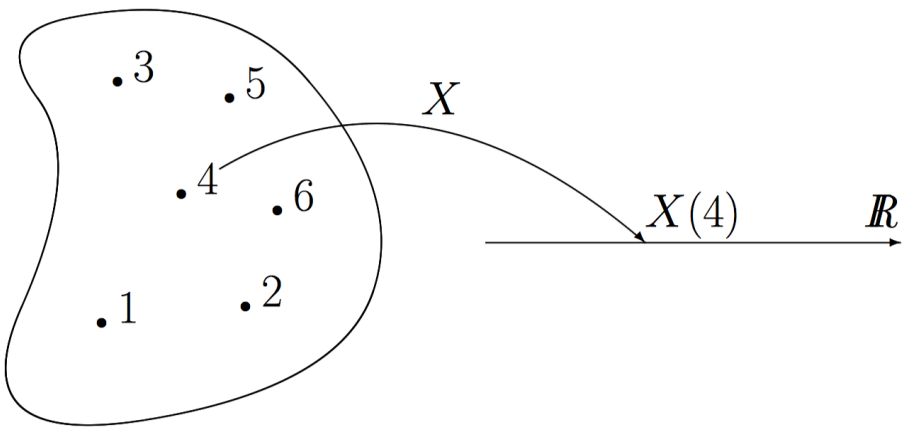
- Definition (**probability**)
  - P defined on a measurable space $(\Omega, \mathcal{A})$ is **a set function** $P : \mathcal{A} \to [0, 1]$ such that (probability axioms).
    1. $P(\varnothing) = 0$
    2. $P(A) \geq 0 \ \forall A \subseteq \Omega$
    3. For disjoint sets A_i and $A_j \Rightarrow P(\cup_{i=1}^{k} A_i) = \sum_{i=1}^{k} P(A_i)$ (countable additivity)
    4. $P(\Omega) = 1$

Random variable

Random variable

- **random variable**:
  A random variable is a real-valued function defined on Ω that is measurable w.r.t. the probability space $(\Omega, \mathcal{A}, P)$ and the Borel measurable space $(\mathbb{R}, \mathcal{B})$, i.e.,

$$X : \Omega \to \mathbb{R} \text{ such that } \forall B \in \mathcal{B}, X^{-1}(B) \in \mathcal{A}.$$



  - What is random here?
  - What is the result of carrying out the random experiment?

Random variable

- **discrete random variable**: There is a discrete set $\{x_i : i = 1, 2, \cdots\}$ such that $\sum P(X = x_i) = 1$.
- **probability mass function**: $p_X(x) \triangleq P(X = x)$ that satisfies
  1. $0 \leq p_X(x) \leq 1$
  2. $\sum_x p_X(x) = 1$
  3. $P(X \in B) = \sum_{x \in B} p_X(x)$

Random variable

- **continuous random variable**

  There is an integrable function $f_X(x)$ such that $P(X \in B) = \int_B f_X(x)dx$.

- **probability density function**

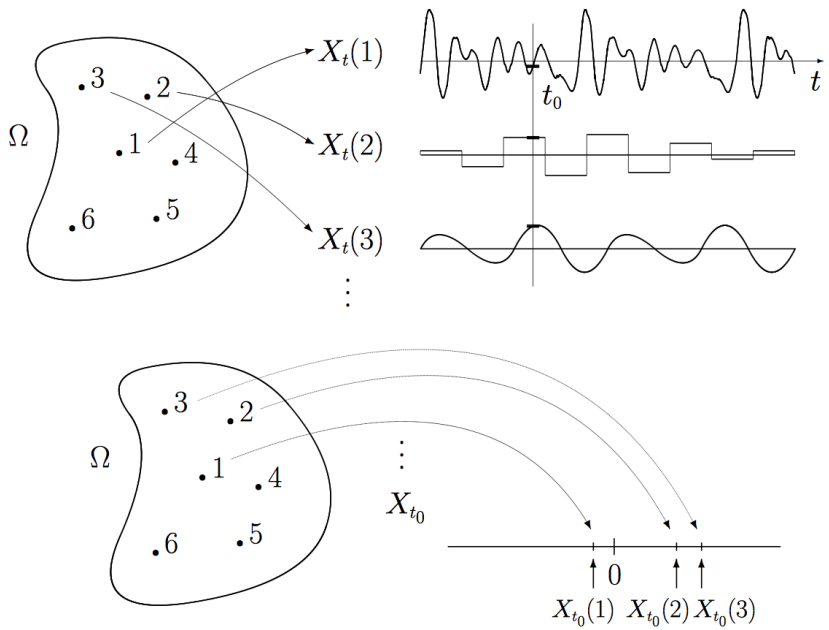
  $f_X(x) \triangleq \lim_{\Delta x \to 0} \frac{P(x < X \leq x + \Delta x)}{\Delta x}$ that satisfies

  1. $f_X(x) > 1$ is possible.
  2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$
  3. $P(X \in B) = \int_{x \in B} f_X(x)dx$

Random process

Random process

- **random process** $X_t(w), t \in I$:
  1. random sequence, random function, or random signal:
     $X_t : \Omega \to$ the set of all sequences or functions
  2. indexed family of infinite number of random variables:
     $X_t : I \to$ set of all random variables defined on Ω
  3. $X_t : \Omega \times I \to \mathbb{R}$
  4. If t is fixed, then a random process becomes a random variable.



Random process

- A random process X_t is completely characterized if the following is known.
  - $P((X_{t_1}, \cdots, X_{t_k}) \in B)$ for any B, k, and $t_1, \cdots, t_k$
- Note that given a random process, only 'finite-dimensional' probabilities or probability functions can be specified.

Functional analysis

Functional analysis

- Definition (**Hilbert space**)
  - Inner product space containing Cauchy sequence limits.
    $\Rightarrow$ Complete space
    $\Rightarrow$ Always possible to *fill all the holes*.
    $\Rightarrow$ $\mathbb{R}$ is complete, $\mathbb{Q}$ is not complete.

Functional analysis

- Definition (**Kernel**)
  - Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,
$$k(x, x') \triangleq \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}.$$

- Note that there is almost no condition on $\mathcal{X}$.

Functional analysis

- Theorem (**Mercer**)
  - Let $(\mathcal{X}, \mu)$ be a finite measurable space and $k \in L_\infty(\mathcal{X}^2, \mu^2)$ be a kernel such that $T_k : L_2(\mathcal{X}, \mu) \to L_2(\mathcal{X}, \mu)$ is positive definite.
  - Let $\phi_i \in L_2(\mathcal{X}, \mu)$ be the normalized eigenfunctions of T_k associated with the eigenvalues $\lambda_i > 0$. Then:
    1. The eigenvalues $\{\lambda_i\}_{i=1}^{\infty}|$ are absolutely summable.
    2. $$k(x, x') = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(x')$$
       holds μ^2 almost everywhere, where the series converges absolutely and uniformly μ^2 almost everywhere.
- Absolutely summable is more important than it seems.
- SB: Mercer's theorem can be interpreted as an infinite dimensional SVD.

Functional analysis

- Definition (**reproducing kernel Hilbert space**)
  - Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}$-valued functions on $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel on $\mathcal{H}$, and $\mathcal{H}$ is a reproducing kernel Hilbert space if
    1. $\forall x \in \mathcal{X}$
       $k(\cdot, x) \in \mathcal{H}$
    2. $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}$
       $\langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ (reproducing property)
    3. $\forall x, x' \in \mathcal{X}$
       $k(x, x') = \langle k(\cdot, x), k(\cdot, x') \rangle_{\mathcal{H}}$
- What does this indicates?

Functional analysis

- Suppose we have a RKHS $\mathcal{H}$, $f(\cdot) \in \mathcal{H}$, and $k(\cdot, x) \in \mathcal{H}$.
- Then the reproducing property indicates that evaluation of $f(\cdot)$ at x, i.e., $f(x)$ is the inner-product of $k(\cdot, x)$ and $f(\cdot)$ itself, i.e.,

$$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}.$$

- Recall Mercer's theorem $k(x, x') = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(x')$. Then,

$$\begin{aligned} f(x) &= \left\langle f, \sum_{i=1}^{\infty} \lambda_i \phi_i(\cdot) \phi_i(x) \right\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^{\infty} \lambda_i \langle f, \phi_i(\cdot) \rangle_{\mathcal{H}} \phi_i(x) \\ &= \sum_{i=1}^{\infty} \bar{\lambda}_i \phi_i(x) \end{aligned}$$

where $\bar{\lambda}_i = \lambda_i \langle f, \phi_i(\cdot) \rangle_{\mathcal{H}}$.

Gaussian process

Gaussian process

- **Gaussian process**: A random process $X(t)$ is a **Gaussian process** if for all $k \in \mathbb{N}$ for all $t_1, \ldots, t_k$, a random vector formed by $X(1), \ldots, X(t_k)$ is jointly Gaussian.
- The joint density is completely specified by
  - Mean: $m(t) = \mathbb{E}(X(t))$, where $m(\cdot)$ is known as a mean function.
  - Covariance: $k(t, s) = \mathbf{cov}(X(t), X(s)) =$, where $k(\cdot, \cdot)$ is known as a covariance function.
- Notation: $X(t) \sim \mathcal{GP}(m(t), k(t, s))$

Function space view

Let $f(\mathbf{x})$ be a (zero-mean) Gaussian process. Then $f(\mathbf{X}) = \begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \in \mathbb{R}^n$

and $f(\mathbf{x}_*) \in \mathbb{R}$ are jointly Gaussian, i.e.,

$$\begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_n) \\ f(\mathbf{x}_*) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_n) & k(\mathbf{x}_1, \mathbf{x}_*) \\ \vdots & \ddots & \vdots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \cdots & k(\mathbf{x}_n, \mathbf{x}_n) & k(\mathbf{x}_n, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{x}_1) & \cdots & k(\mathbf{x}_*, \mathbf{x}_n) & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right).$$

Function space view

- We rewrite the joint Gaussian distribution as

$$\begin{bmatrix} \mathbf{f} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) & K(X,\mathbf{x}_*) \\ K(\mathbf{x}_*,X) & k(\mathbf{x}_*,\mathbf{x}_*) \end{bmatrix}\right).$$

- Recall that the conditional distribution $p(\mathbf{x}|\mathbf{y})$ of a jointly Gaussian random vector $\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$ is also a Gaussian random vector with mean $\mathbb{E}(\mathbf{x}|\mathbf{y})$ and covariance matrix $\Sigma_{\mathbf{x}|\mathbf{y}}$ where

$$\mathbb{E}(\mathbf{x}|\mathbf{y}) = \mathbb{E}(\mathbf{x}) + \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1}(\mathbf{y} - \mathbb{E}(\mathbf{y}))$$
$$\Sigma_{\mathbf{x}|\mathbf{y}} = \Sigma_{\mathbf{xx}} - \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}}.$$

Function space view

By conditioning, we get

$$f_*|\mathbf{x}_*, X, \mathbf{f} \sim \mathcal{N}(\mu_*, \sigma_*^2)$$

where

$$\mu_* = K(\mathbf{x}_*, X)K(X, X)^{-1}\mathbf{f}$$

and

$$\sigma_* = k(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, X)K(X, X)^{-1}K(X, \mathbf{x}_*).$$

Function space view

- In the previous case, measurement noise is not included.
- Let $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$. Then the covariance between two outputs becomes:

$$\mathbf{cov}(y(\mathbf{x}_1, \mathbf{x}_2)) = k(\mathbf{x}_1, \mathbf{x}_2) + \sigma_n^2.$$

- Consequently, the joint distribution of y and $\mathbf{f}_*$ becomes:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X, \mathbf{x}_*) \\ K(\mathbf{x}_*, X) & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right).$$

Function space view

By conditioning, we get

$$f_*|\mathbf{x}_*, X, \mathbf{y} \sim \mathcal{N}(\mu_*, \sigma_*^2)$$

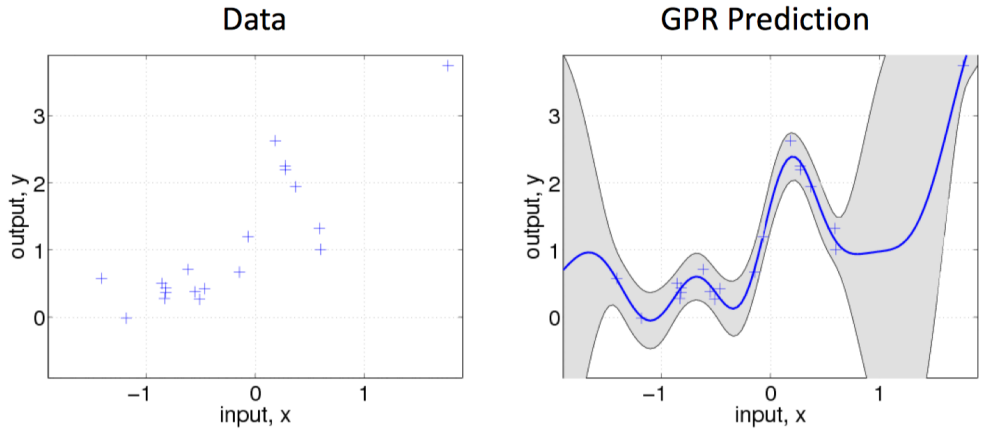
where

$$\mu_* = K(\mathbf{x}_*, X)(K(X, X) + \sigma_n^2 I)^{-1}\mathbf{y}$$

and

$$\sigma_* = k(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, X)(K(X, X) + \sigma_n^2 I)^{-1}K(X, \mathbf{x}_*).$$

Comments on Gaussian process regression



- **Pros**: principled, probabilistic, predictive uncertainty
- **Cons**: computationally intensive ($O(n^3)$ where n is the number of data)

Summary of Variational Inference

Summary of Variational Inference

- Log marginal likelihood
  $=$ ELBO (variational free energy) $+ D_{KL}(q(w|\theta)||p(w|D))$
- Note that $p(w|D)$ can hardly be computed analytically.

Summary of Variational Inference

- Derivation:

$$
\begin{aligned}
\ln p(D) &= \int \ln p(D) q(w|\theta) dw \\
&= \int q(w|\theta) \ln \frac{p(D)p(w|D)}{p(w|D)} dw \\
&= \int q(w|\theta) \ln \frac{p(w,D)}{p(w|D)} dw \\
&= \int q(w|\theta) \ln \frac{p(D|w)p(w)}{p(w|D)} dw \\
&= \int q(w|\theta) \ln \frac{q(w|\theta)p(D|w)p(w)}{q(w|\theta)p(w|D)} dw \\
&= \int q(w|\theta) \ln \frac{q(w|\theta)}{p(w|D)} dw + \int q(w|\theta) \ln \frac{p(D|w)p(w)}{q(w|\theta)} dw \\
&= D_{KL}(q(w|\theta)||p(w|D)) + \mathcal{F}[q]
\end{aligned}
$$

where $\mathcal{F}[q]$ is the variational free energy or ELBO.

Summary of Variational Inference

- The variational free energy or ELBO can further experessed as:
- Derivation:

$$\begin{aligned}
\mathcal{F}[q] &= \int q(w|\theta) \ln \frac{p(D|w)p(w)}{q(w|\theta)} dw \\
&= \int q(w|\theta) \ln p(D|w) dw + \int q(w|\theta) \ln \frac{p(w)}{q(w|\theta)} dw \\
&= \mathbb{E}_{q(w|\theta)}[\ln p(D|w)] - D_{KL}(q(w|\theta)||p(w)) \\
&= \text{likelihood under } q - \text{prior fitting term}
\end{aligned}$$

- We try to maximize $\mathcal{F}[q]$ to

1. maximize the ~~marginal~~ likelihood $p(D)$
2. ~~reduce~~ the gap between $p(w|D)$ and $q(w)$.
3. keep the variational distribution $Q(w)$ close to our prior $p(w|A)$.

Stein Variational Gradient Descent

Stein Variational Gradient Descent

- **Stein Variational Gradient Descent**
  - To implement the iterative procedure, (unnormalized) posterior $p(x)$ is given.
  - Then, we draw a set of particles $\{x_i^0\}_{i=1}^n$ for the initial distribution q_0.
  - Each particle x_i is updated as follows:

$$x_i^{l+1} \leftarrow x_i^l + \epsilon_l \hat{\phi}^*(x_i^l)$$

where

$$\hat{\phi}^*(x) = \frac{1}{n} \sum_{j=1}^{n} \left[k(x_j^l, x) \nabla_{x_j^l} \log p(x_j^l) + \nabla_{x_j^l} k(x_j^l, x) \right].$$

Stein Variational Gradient Descent

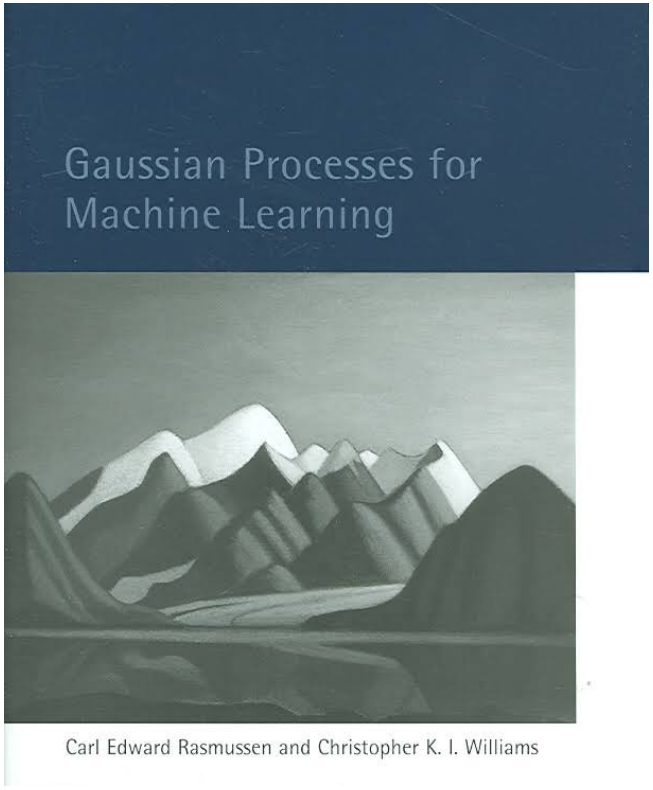
- The update rule has nice interpretations:

$$\hat{\phi}^*(x) = \frac{1}{n}\sum_{j=1}^{n}\Big[k(x_j^l, x)\nabla_{x_j^l}\log p(x_j^l) + \nabla_{x_j^l}k(x_j^l, x)\Big].$$

1. $k(x_j^l, x)$: Similarity between current particle to update x and j-th particle x_j^l.
2. $\nabla_{x_j^l}\log p(x_j^l)$: Particle update direction of current particle to update x where it is computed from j-th particle x_j^l.
   - Note that as we are using the score function, i.e., $\nabla_x \log p(x)$, unnormalized $p(x)$ can be used!
   - This is also used in policy gradient methods such as REINFORCE.
3. $\nabla_{x_j^l}k(x_j^l, x)$: This term can be interpreted as running a gradient ascent method on $k(\cdot,\cdot)$. As the value of a kernel function usually increases as the distance between two inputs decreases, it can be interpreted as an attractive force between particles.

Questions?

Must read



Gaussian process for machine learning [1]

References i


C. E. Rasmussen and C. K. Williams.
Gaussian processes for machine learning, volume 1.
MIT press Cambridge, 2006.