Mathematics for Machine Learning

— Probabilistic Modeling & Inference

Joseph Chuang-Chieh Lin

Department of Computer Science & Information Engineering,
Tamkang University

Fall 2023

Credits for the resource

- The slides are based on the textbooks:
  - *Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong: Mathematics for Machine Learning. Cambridge University Press. 2020.*
  - *Howard Anton, Chris Rorres, Anton Kaul: Elementary Linear Algebra. Wiley. 2019.*
- We could partially refer to the monograph:
  Francesco Orabona: A Modern Introduction to Online Learning. https://arxiv.org/abs/1912.13213

Outline

1. Probabilistic Models & Bayesian Inference

2. Latent-Variable Models

Motivation

- We are concerned with prediction of future events and decision making.
- We build models that describe the generative process that generates the observed data.
- For example, consider the outcome of a coin-flip experiment ("heads" or "tails").

Motivation

- We are concerned with prediction of future events and decision making.
- We build models that describe the generative process that generates the observed data.
- For example, consider the outcome of a coin-flip experiment ("heads" or "tails").
    - Define a parameter μ which describes the probability of "heads" (the parameter of a Bernoulli distribution).

Motivation

- We are concerned with prediction of future events and decision making.
- We build models that describe the generative process that generates the observed data.
- For example, consider the outcome of a coin-flip experiment ("heads" or "tails").
  - Define a parameter μ which describes the probability of "heads" (the parameter of a Bernoulli distribution).
  - Then, we can sample an outcome $x \in \{\text{head}, \text{tail}\}$ from the Bernoulli distribution $p(x \mid \mu) = \text{Ber}(\mu)$.

Motivation

- We are concerned with prediction of future events and decision making.
- We build models that describe the generative process that generates the observed data.
- For example, consider the outcome of a coin-flip experiment ("heads" or "tails").
  - Define a parameter μ which describes the probability of "heads" (the parameter of a Bernoulli distribution).
  - Then, we can sample an outcome $x \in \{\text{head}, \text{tail}\}$ from the Bernoulli distribution $p(x \mid \mu) = \text{Ber}(\mu)$.
- **Note:** μ is unknown in advance and can never be observed directly.

Motivation

- We are concerned with prediction of future events and decision making.
- We build models that describe the generative process that generates the observed data.
- For example, consider the outcome of a coin-flip experiment ("heads" or "tails").
  - Define a parameter μ which describes the probability of "heads" (the parameter of a Bernoulli distribution).
  - Then, we can sample an outcome $x \in \{\text{head}, \text{tail}\}$ from the Bernoulli distribution $p(x \mid \mu) = \text{Ber}(\mu)$.
- **Note:** μ is unknown in advance and can never be observed directly.
- We need mechanisms to learn something about μ given observed outcomes of coin-flip.

Probabilistic Models

- The benefit of using probabilistic models:
  - A unified and consistent set of tools from probability theory for modeling, inference, prediction, and model selection.
- $p(\mathbf{x}, \boldsymbol{\theta})$: the joint distribution of the observed variables $\mathbf{x}$ and the hidden parameters $\boldsymbol{\theta}$.

Probabilistic Models

- The benefit of using probabilistic models:
  - A unified and consistent set of tools from probability theory for modeling, inference, prediction, and model selection.
- $p(\mathbf{x}, \boldsymbol{\theta})$: the joint distribution of the observed variables $\mathbf{x}$ and the hidden parameters $\boldsymbol{\theta}$. It encapsulates the information:
  - The prior and the likelihood.
  - The marginal likelihood $p(\mathbf{x})$ (though integrating out the parameters is required.)
  - The posterior (obtained by dividing the joint by the marginal likelihood).

Probabilistic Models

- The benefit of using probabilistic models:
  - A unified and consistent set of tools from probability theory for modeling, inference, prediction, and model selection.
- $p(\mathbf{x}, \boldsymbol{\theta})$: the joint distribution of the observed variables $\mathbf{x}$ and the hidden parameters $\boldsymbol{\theta}$. It encapsulates the information:
  - The prior and the likelihood.
  - The marginal likelihood $p(\mathbf{x})$ (though integrating out the parameters is required.)
  - The posterior (obtained by dividing the joint by the marginal likelihood).
- Therefore, a probabilistic model is specified by the joint distribution of all its random variables.

Bayesian Inference (1/3)

- We have already learnt two ways of estimating model parameters $\boldsymbol{\theta}$:
  - Maximum likelihood estimation (MLE)
  - Maximum a posteriori estimation (MAP)
- We can then obtain a *single-best* value of $\boldsymbol{\theta}$ (solving an optimization problem), then we can use them to make predictions.
- Note: These decision-making systems typically have different objective functions than the likelihood (e.g., squared-error loss or a mis-classification error).
- Having the full posterior distribution around can be useful and leads to more robust decisions.

Bayesian Inference (2/3)

- Bayesian inference: finding such a posterior distribution.
- For a dataset $\mathcal{X}$, a parameter prior $p(\boldsymbol{\theta})$, and a likelihood function, the posterior

$$p(\boldsymbol{\theta} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{X})},$$

then by applying Bayes' theorem,

$$p(\mathcal{X}) = \int p(\mathcal{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$

Bayesian Inference (2/3)

- Bayesian inference: finding such a posterior distribution.
- For a dataset $\mathcal{X}$, a parameter prior $p(\boldsymbol{\theta})$, and a likelihood function, the posterior
$$p(\boldsymbol{\theta} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{X})},$$
then by applying Bayes' theorem,
$$p(\mathcal{X}) = \int p(\mathcal{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$
- Propagate uncertainty from the parameters to the data. Specifically, with a distribution $p(\boldsymbol{\theta})$, our predictions will be
$$p(\mathbf{x}) = \int p(\mathbf{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$$

Bayesian Inference (2/3)

- Bayesian inference: finding such a posterior distribution.
- For a dataset $\mathcal{X}$, a parameter prior $p(\boldsymbol{\theta})$, and a likelihood function, the posterior

$$p(\boldsymbol{\theta} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{X})},$$

  then by applying Bayes' theorem,

$$p(\mathcal{X}) = \int p(\mathcal{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$

- Propagate uncertainty from the parameters to the data. Specifically, with a distribution $p(\boldsymbol{\theta})$, our predictions will be

$$p(\mathbf{x}) = \int p(\mathbf{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\theta}}[p(\mathbf{x} \mid \boldsymbol{\theta})],$$

  which no longer depend on the model parameters $\boldsymbol{\theta}$.

Bayesian Inference (2/3)

- Bayesian inference: finding such a posterior distribution.
- For a dataset $\mathcal{X}$, a parameter prior $p(\boldsymbol{\theta})$, and a likelihood function, the posterior

$$p(\boldsymbol{\theta} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{X})},$$

then by applying Bayes' theorem,

$$p(\mathcal{X}) = \int p(\mathcal{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$

- Propagate uncertainty from the parameters to the data. Specifically, with a distribution $p(\boldsymbol{\theta})$, our predictions will be

$$p(\mathbf{x}) = \int p(\mathbf{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\theta}}[p(\mathbf{x} \mid \boldsymbol{\theta})],$$

which no longer depend on the model parameters $\boldsymbol{\theta}$.

  - It has been marginalized/integrated out.

Bayesian Inference (3/3)

$$p(\mathbf{x}) = \int p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\theta}}[p(\mathbf{x} \mid \boldsymbol{\theta})],$$

- The prediction becomes an average over all plausible parameter values $\boldsymbol{\theta}$.
  - The plausibility is encapsulated by the distribution $p(\boldsymbol{\theta})$.

Computational Issues

- MLE or MAP yields a consistent point estimate $\boldsymbol{\theta}^*$ of the parameters.
  - Key computational problem: optimization.
  - Prediction: straightforward.
- Bayesian inference yields a distribution.
  - Key computational problem: integration.
  - Prediction: solving another integration problem.

Outline

Latent Variables

- Sometimes it is useful to have additional variable (besides $\boldsymbol{\theta}$) as part of the model.
    - We call them latent variables.
- Latent variables can
    - Describe the data-generation process.
    - Increase the interpretability of the model.
    - Simplify the structure of the model.

In the Data Generation Process

Denote data by $\mathbf{x}$, the model parameter by $\boldsymbol{\theta}$ and the latent variables by $\mathbf{z}$, we obtain the conditional distribution:

$$p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta}).$$

In the Data Generation Process

Denote data by $\mathbf{x}$, the model parameter by $\boldsymbol{\theta}$ and the latent variables by $\mathbf{z}$, we obtain the conditional distribution:

$$p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta}).$$

This allows us to generate data for any model parameter and latent variables.

In the Data Generation Process

Denote data by $\mathbf{x}$, the model parameter by $\boldsymbol{\theta}$ and the latent variables by $\mathbf{z}$, we obtain the conditional distribution:

$$p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta}).$$

This allows us to generate data for any model parameter and latent variables.

- We place a prior $p(\mathbf{z})$ on the given latent variables $\mathbf{z}$.

In the Data Generation Process

Denote data by $\mathbf{x}$, the model parameter by $\boldsymbol{\theta}$ and the latent variables by $\mathbf{z}$, we obtain the conditional distribution:

$$p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta}).$$

This allows us to generate data for any model parameter and latent variables.

- We place a prior $p(\mathbf{z})$ on the given latent variables $\mathbf{z}$.

A Two-Step Procedure

1. Compute the likelihood $p(\mathbf{x} \mid \boldsymbol{\theta})$ (not depending on $\mathbf{z}$).
2. Use the likelihood for parameter estimation or Bayesian inference.

Likelihood in Terms of Marginal Distribution

What we already have: a conditional distribution

$$p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta}).$$

We need to marginalize out the latent variables to have the predictive distribution of the data given the model parameters $\boldsymbol{\theta}$:

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \int p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}) \mathrm{d}\mathbf{z},$$

Likelihood in Terms of Marginal Distribution

What we already have: a conditional distribution

$$p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta}).$$

We need to marginalize out the latent variables to have the predictive distribution of the data given the model parameters $\boldsymbol{\theta}$:

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \int p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}) \mathrm{d}\mathbf{z},$$

Note that $p(\mathbf{z})$ is a prior,

Likelihood in Terms of Marginal Distribution

What we already have: a conditional distribution

$$p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta}).$$

We need to marginalize out the latent variables to have the predictive distribution of the data given the model parameters $\boldsymbol{\theta}$:

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \int p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}) \mathrm{d}\mathbf{z},$$

Note that $p(\mathbf{z})$ is a prior, and $(\mathbf{x} \mid \boldsymbol{\theta})$ does not depend on $\mathbf{z}$.

Bayesian Inference in a Latent Variable Model

Place a prior $p(\boldsymbol{\theta})$ and use Bayes' theorem to obtain

Bayesian Inference in a Latent Variable Model

Place a prior $p(\boldsymbol{\theta})$ and use Bayes' theorem to obtain

$$p(\boldsymbol{\theta} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathcal{X})}$$

Bayesian Inference in a Latent Variable Model

Place a prior $p(\boldsymbol{\theta})$ and use Bayes' theorem to obtain

$$p(\boldsymbol{\theta} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{X})}$$

$\Rightarrow$ a posterior distribution over the model parameters given a dataset $\mathcal{X}$.

Bayesian Inference in a Latent Variable Model

Place a prior $p(\boldsymbol{\theta})$ and use Bayes' theorem to obtain

$$p(\boldsymbol{\theta} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{X})}$$

$\Rightarrow$ a posterior distribution over the model parameters given a dataset $\mathcal{X}$.

- $p(\mathcal{X} \mid \boldsymbol{\theta})$ requires the marginalization of latent variables $\mathbf{z}$.

A Posterior on the Latent Variables

$$p(\mathbf{z} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \mathbf{z})p(\mathbf{z})}{p(\mathcal{X})}, \quad p(\mathcal{X} \mid \mathbf{z}) = \int p(\mathcal{X} \mid \mathbf{z}, \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},$$

A Posterior on the Latent Variables

$$p(\mathbf{z} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \mathbf{z})p(\mathbf{z})}{p(\mathcal{X})}, \quad p(\mathcal{X} \mid \mathbf{z}) = \int p(\mathcal{X} \mid \mathbf{z}, \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},$$

where $p(\mathbf{z})$ is the prior on $\mathbf{z}$, and $p(\mathcal{X} \mid \mathbf{z})$ requires us to integrate out hte model parameters $\boldsymbol{\theta}$.

A Posterior on the Latent Variables

$$p(\mathbf{z} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \mathbf{z})p(\mathbf{z})}{p(\mathcal{X})}, \quad p(\mathcal{X} \mid \mathbf{z}) = \int p(\mathcal{X} \mid \mathbf{z}, \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},$$

where $p(\mathbf{z})$ is the prior on $\mathbf{z}$, and $p(\mathcal{X} \mid \mathbf{z})$ requires us to integrate out hte model parameters $\boldsymbol{\theta}$.

- Note that it may be difficult to solve the integrals analytically.

A Posterior on the Latent Variables

$$p(\mathbf{z} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \mathbf{z})p(\mathbf{z})}{p(\mathcal{X})}, \quad p(\mathcal{X} \mid \mathbf{z}) = \int p(\mathcal{X} \mid \mathbf{z}, \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},$$

where $p(\mathbf{z})$ is the prior on $\mathbf{z}$, and $p(\mathcal{X} \mid \mathbf{z})$ requires us to integrate out hte model parameters $\boldsymbol{\theta}$.

- Note that it may be difficult to solve the integrals analytically.
- An easier quantity to compute:

A Posterior on the Latent Variables

$$p(\mathbf{z} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \mathbf{z})p(\mathbf{z})}{p(\mathcal{X})}, \quad p(\mathcal{X} \mid \mathbf{z}) = \int p(\mathcal{X} \mid \mathbf{z}, \boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta},$$

where $p(\mathbf{z})$ is the prior on $\mathbf{z}$, and $p(\mathcal{X} \mid \mathbf{z})$ requires us to integrate out hte model parameters $\boldsymbol{\theta}$.

- Note that it may be difficult to solve the integrals analytically.
- An easier quantity to compute:

$$p(\mathbf{z} \mid \mathcal{X}, \boldsymbol{\theta}) = \frac{p(\mathcal{X} \mid \mathbf{z}, \boldsymbol{\theta})p(\mathbf{z})}{p(\mathcal{X} \mid \boldsymbol{\theta})},$$

  - $p(\mathbf{z})$: the prior on $\mathbf{z}$; $p(\mathcal{X} \mid \mathbf{z}, \boldsymbol{\theta})$: given.

Example

Consider the set of affine functions.

- Let $\mathbf{x}_i = [1, x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(D)}]^\top$
- The corresponding parameter $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots, \theta_D]^\top$.
- Consider a more compact form as below:

$$f(\mathbf{x}_i, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}_i.$$

Example

Consider the set of affine functions.

- Let $\mathbf{x}_i = [1, x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(D)}]^\top$
- The corresponding parameter $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots, \theta_D]^\top$.
- Consider a more compact form as below:

$$f(\mathbf{x}_i, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}_i.$$

which is equivalent to

$$f(\mathbf{x}_i, \boldsymbol{\theta}) = \theta_0 + \sum_{d=1}^{D} \theta_d x_i^{(d)}$$

Discussions