Intro to Bayesian Statistics

Maxim Panov

Skoltech

November 2021

Skoltech

Outline

Outline

Frequentist (Classical) Point of View

F1 Probability refers to limiting relative frequencies.

- Probabilities are objective properties of the real world.

F2 Parameters are fixed, unknown constants.

- Because they are not fluctuating, no useful probability statements can be made about parameters.

F3 Statistical procedures should be designed to have well-defined long run frequency properties.

- For example, a 95 percent confidence interval should trap the true value of the parameter with limiting frequency at least 95 percent.

Bayesian Philosophy

B1 Probability describes degree of belief, not limiting frequency.

- For example, I might say that "the probability that Albert Einstein drank a cup of tea on August 1, 1948" is 0.35.
- This does not refer to any limiting frequency.
- It reflects my strength of belief that the proposition is true.

B2 We can make probability statements about parameters, even though they are fixed constants.

B3 We make inferences about a parameter θ by producing a probability distribution for θ.

- Inferences, such as point estimates and interval estimates, may then be extracted from this distribution.

Bayesian Approach to Machine Learning

Consider a probabilistic model:

$$p(y \mid \mathbf{x}, \mathbf{w}),$$

where

- $\mathbf{x}$ is a model input;
- $\mathbf{w}$ is a vector of model parameters (i.e., linear regression weights).

Let us be given the dataset $\mathcal{D}_n = \{\mathbf{x}_i, y_i\}_{i=1}^n$. Then the likelihood of the data reads as

$$p(\mathcal{D}_n \mid \mathbf{w}) = \prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i, \mathbf{w}).$$

In Bayesian approach, $\mathbf{w}$ is assumed to be a random variable with some prior distribution:

$$\mathbf{w} \sim p(\mathbf{w}).$$

Bayesian Inference problem

In Bayesian problems we are interested in posterior distribution of latent variables:

$$p(\mathbf{w} \mid \mathcal{D}_n) = \frac{p(\mathcal{D}_n \mid \mathbf{w})\, p(\mathbf{w})}{p(\mathcal{D}_n)},$$

where $\mathcal{D}_n$ – observed data, $\mathbf{w}$ – latent (unobserved) variables.

Posterior allows **to reason about the uncertainties** in latent variables.

The following distributions are involved:

- $p(\mathbf{w} \mid \mathcal{D}_n)$ – posterior (our updated knowledge about $\mathbf{w}$ after we have observed data $\mathcal{D}_n$);
- $p(\mathbf{w})$ – prior (our knowledge about $\mathbf{w}$ before we have observed data $\mathcal{D}_n$);
- $p(\mathcal{D}_n \mid \mathbf{w})$ – likelihood (probability of data $\mathcal{D}_n$ given latent variables $\mathbf{w}$);
- $p(\mathcal{D}_n)$ – normalizing constant for $p(\mathbf{w} \mid \mathcal{D}_n)$ to be a proper distribution.

Bayesian vs. Frequentist

- MLE:

$$\widehat{\mathbf{w}} = \arg\max_{\mathbf{w}} \log p(\mathcal{D}_n \mid \mathbf{w}).$$

- Maximum a posteriori estimate (MAP).
  - Posterior:

$$p(\mathbf{w} \mid \mathcal{D}_n) = \frac{p(\mathcal{D}_n \mid \mathbf{w})p(\mathbf{w})}{p(\mathcal{D}_n)}.$$

  - MAP:

$$\widehat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{w} \mid \mathcal{D}_n).$$

  - MAP $\equiv$ regularized MLE:

$$\widehat{\mathbf{w}} = \arg\max_{\mathbf{w}} [\log p(\mathcal{D}_n \mid \mathbf{w}) + \log p(\mathbf{w})].$$

Full Bayesian Approach

Let us be given the dataset $\mathcal{D}_n = \{\mathbf{x}_i, y_i\}_{i=1}^n$.

We can compute a posterior distribution:

$$p(\mathbf{w} \mid \mathcal{D}_n) = \frac{p(\mathcal{D}_n \mid \mathbf{w})p(\mathbf{w})}{\int p(\mathcal{D}_n \mid \mathbf{w})p(\mathbf{w})d\mathbf{w}}.$$

Posterior predictive distribution:

$$p(y \mid \mathbf{x}, \mathcal{D}_n) = \int p(y \mid \mathbf{x}, \mathbf{w})\, p(\mathbf{w} \mid \mathcal{D}_n)\, d\mathbf{w} = \mathbb{E}_{p(\mathbf{w} \mid \mathcal{D}_n)}\, p(y \mid \mathbf{x}, \mathbf{w}).$$

- $\mathbf{w}$ is integrated out.
- To compute full posterior no optimization wrt. $\mathbf{w}$ is needed!
- Such an approach is sometimes called Full Bayesian Approach.

Bernstein – von Mises Theorem

Maximum Likelihood Estimation (MLE)

$$\widehat{\boldsymbol{\theta}}_n = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}).$$

$\widehat{\boldsymbol{\theta}}$ concentrates around the "true" parameter $\boldsymbol{\theta}_*$ (Fisher Theorem):

$$D(\boldsymbol{\theta}_*)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*) \xrightarrow{P} \mathcal{N}(0, I_p), \quad n \to \infty,$$

where $D^2(\boldsymbol{\theta}_*)$ is the Fisher information matrix at $\boldsymbol{\theta}_*$, $p = \dim(\boldsymbol{\theta})$.

Bernstein – von Mises Theorem

$$D(\boldsymbol{\theta}_*)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n) \mid \mathcal{D}_n \xrightarrow{TV} \mathcal{N}(0, I_p), \quad n \to \infty.$$

Features:

– concentration of posterior, near Gaussianity;

– the limit doesn't depend on $p(\boldsymbol{\theta})$.

History of BvM: S. Bernstein

Sergey N. Bernstein
(1917, lecture notes published by Kharkiv university)

- Consider Bernoulli model with probability of success p.
- Let n experiments were carried out with m successes.
- Let there exists $\varepsilon > 0$ such that the prior density $\pi(p)$ is different from 0 for $p \in (\frac{m}{n} - \varepsilon, \frac{m}{n} + \varepsilon)$.
- Then $\mathbb{P}\big(p \in (x, x + \Delta x) \mid m\big) \to \Phi\big((x, x + \Delta x)\big)$, $n \to \infty$.

Skoltech
Skolkovo Institute of Science and Technology

History of BvM: R. von Mises

Richard von Mises
(1931, "Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik")

- Consider multinomial distribution with k possible outcomes and their probabilities $p_1, \ldots, p_k$.
- Let n experiments were carried out with number of outcomes of each kind $m_1, \ldots, m_k$.
- Some technical conditions on the prior $\pi(p_1, \ldots, p_k)$ are satisfied.
- Consider twice continuously differentiable function $f(x_1, \ldots, x_k)$.
- If $n \to \infty$ then $\mathbb{P}\big(f(p_1, \ldots, p_k) \in (x, x + \Delta x) \mid m_1, \ldots, m_k\big) \to \Phi((x, x + \Delta x))$.

Skoltech
Skolkovo Institute of Science and Technology

The Exponential Family of Distributions

Exponential family:

$$p(\mathbf{x}) = h(\mathbf{x}) \cdot e^{\boldsymbol{\theta}^{\mathrm{T}} T(\mathbf{x}) - A(\boldsymbol{\theta})},$$

where

- $\boldsymbol{\theta}$ — vector of parameters,
- $T(\mathbf{x})$ — vector of sufficient statistics,
- $A(\boldsymbol{\theta})$ — cumulant generating function.

Key point: $\mathbf{x}$ and $\boldsymbol{\theta}$ only “mix” in $e^{\boldsymbol{\theta}^{\mathrm{T}} T(\mathbf{x})}$.

The Exponential Family of Distributions

Exponential family:

$$p(\mathbf{x}) = h(\mathbf{x}) \cdot e^{\boldsymbol{\theta}^{\mathrm{T}} T(\mathbf{x}) - A(\boldsymbol{\theta})}.$$

To get a normalized distribution for any $\boldsymbol{\theta}$

$$\int p(\mathbf{x}) d\mathbf{x} = e^{-A(\boldsymbol{\theta})} \int h(\mathbf{x}) e^{\boldsymbol{\theta}^{\mathrm{T}} T(\mathbf{x})} d\mathbf{x} = 1$$

so

$$e^{A(\boldsymbol{\theta})} = \int h(\mathbf{x}) e^{\boldsymbol{\theta}^{\mathrm{T}} T(\mathbf{x})} d\mathbf{x}.$$

E.g. for $T(\mathbf{x}) = \mathbf{x}$, $A(\boldsymbol{\theta})$ is the $\log$ of Laplace transform of $h(\mathbf{x})$.

Examples

- Gaussian $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $x \in \mathbb{R}$.
- Bernoulli $p(x) = \alpha^x (1-\alpha)^{1-x}$, $x \in \{0, 1\}$.
- Binomial $p(x) = C_n^x \alpha^x (1-\alpha)^{n-x}$, $x \in \{0, 1, 2, \dots, n\}$.
- Multinomial $p(\mathbf{x}) = \frac{n!}{x_1! x_2! \dots x_n!} \prod_{i=1}^n \alpha_i^{x_i}$, $x_i \in \{0, 1, 2, \dots, n\}$, $\sum_{i=1}^n x_i = n$.
- Exponential $p(x) = \lambda e^{-\lambda x}$, $x \in \mathbb{R}^+$.
- Poisson $p(x) = \frac{e^{-\lambda}}{x!} \lambda^x$, $x \in \{0, 1, 2, \dots\}$.
- Dirichlet $p(\mathbf{x}) = \frac{\Gamma(\sum_{i=1}^n \alpha_i)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^n x_i^{\alpha_i - 1}$, $x_i \in [0, 1]$, $\sum_{i=1}^n x_i = 1$.

Conjugate Priors in Bayesian Statistics

- Posterior distribution:

$$p(\mathbf{w} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{w})p(\mathbf{w})}{\int p(\mathbf{x} \mid \mathbf{w})p(\mathbf{w})d\mathbf{w}}.$$

- Type of a posterior given prior?

$$\underbrace{p(\mathbf{w})}_{parametric} \Rightarrow \underbrace{p(\mathbf{x} \mid \mathbf{w})}_{parametric} \cdot p(\mathbf{w}) \Rightarrow \text{ we get } p(\mathbf{w}|\mathbf{x}) \sim \underbrace{p(\mathbf{x} \mid \mathbf{w}) \cdot p(\mathbf{w})}_{???}.$$

- Conjugacy: require $p(\mathbf{w})$ and $p(\mathbf{w} \mid \mathbf{x})$ to be of the same form. E.g.

$$\underbrace{p(\mathbf{w})}_{Dirichlet} \Rightarrow \underbrace{p(\mathbf{x} \mid \mathbf{w})}_{Multinomial} \cdot p(\mathbf{w}) \Rightarrow \underbrace{p(\mathbf{w} \mid \mathbf{x})}_{Dirichlet}.$$

- $p(\mathbf{w})$ and $p(\mathbf{x} \mid \mathbf{w})$ are then called conjugate distributions.

Example: Dirichlet and Multinomial

$$p(\mathbf{w}) = \frac{\Gamma(\sum_{i=1}^{d} \alpha_i)}{\prod_{i=1}^{d} \Gamma(\alpha_i)} \prod_{i=1}^{d} w_i^{\alpha_i - 1}$$ — Dirichlet in $\mathbf{w}$, $\Gamma(n) = (n-1)!$

$$p(\mathbf{x} \mid \mathbf{w}) = \frac{(\sum_{i=1}^{d} x_i)!}{x_1! x_2! \dots x_d!} \prod_{i=1}^{d} w_i^{x_i}$$ — Multinomial in $\mathbf{x}$;

$$p(\mathbf{w} \mid \mathbf{x}) \sim p(\mathbf{x} \mid \mathbf{w}) p(\mathbf{w}) = \text{const} \times \prod_{i=1}^{d} w_i^{x_i + \alpha_i - 1},$$

which is again Dirichlet, so we must have

$$p(\mathbf{w} \mid \mathbf{x}) = \frac{\Gamma(\sum_{i=1}^{d} \alpha_i + x_i)}{\prod_{i=1}^{d} \Gamma(\alpha_i + x_i)} \prod_{i=1}^{d} w_i^{x_i + \alpha_i - 1}.$$

Conjugate Pairs

- **Prior**: Gaussian $e^{-\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|^2/(2\sigma^2)}$;
  Likelihood: $e^{-\|\mathbf{x}-\boldsymbol{\mu}\|^2/(2\sigma^2)}$.

- **Prior**: Beta $\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} w^{r-1}(1-w)^{s-1}$;
  Likelihood: Bernoulli $w^x(1-w)^{1-x}$.

- **Prior**: Dirichlet $\frac{\Gamma(\sum_{i=1}^d \alpha_i)}{\prod_{i=1}^d \Gamma(\alpha_i)} \prod_{i=1}^d w_i^{\alpha_i-1}$;
  Likelihood: Multinomial $\frac{(\sum_{i=1}^d x_i)!}{\prod_{i=1}^d x_i!} \prod_{i=1}^d w_i^{x_i}$.

Conjugate Pairs

Note: Conjugacy is mutual, e.g. if

$$Dirichlet \Rightarrow Multinomial \Rightarrow Dirichlet$$

then

$$Multinomial \Rightarrow Dirichlet \Rightarrow Multinomial$$

Linear Model

- Linear Basis Function Models

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x})^\top,$$

  where $\boldsymbol{\phi}(\mathbf{x})$ is a vector of known basis functions $\phi_j(\mathbf{x})$
- Typical basis functions

$$\phi_j(\mathbf{x}) = x_{j_1}^{j_0}, \quad \phi_j(\mathbf{x}) = \exp\left\{-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2s^2}\right\},$$

$$\phi(\mathbf{x}) = \sigma\left(\boldsymbol{\mu}_{j,1} \cdot \mathbf{x}^\top + \mu_{j,0}\right), \quad \sigma(a) = \frac{1}{1 + e^{-a}}.$$

- We assume that parameters of basis functions are fixed to some known values.

Maximum Likelihood and Least Squares

- Data model for y (ε is a Gaussian white noise with variance β^{-1}):

$$
\begin{aligned}
y &= f(\mathbf{x}, \mathbf{w}) + \varepsilon, \\
p(y \mid \mathbf{x}, \mathbf{w}, \beta) &= \mathcal{N}(y \mid f(\mathbf{x}, \mathbf{w}), \beta^{-1}),
\end{aligned}
$$

- For $\mathbf{Y}_n = \{y_1, \dots, y_n\}$ and $\mathbf{X}_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ data likelihood

$$
\pi(\mathbf{Y}_n \mid \mathbf{X}_n, \mathbf{w}, \beta) = \prod_{i=1}^{n} \mathcal{N}(y_i \mid \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top, \beta^{-1}).
$$

Maximum Likelihood and Least Squares

- Data log-likelihood has the form

$$\log p(\mathbf{Y}_n \mid \mathbf{X}_n, \mathbf{w}, \beta) = \sum_{i=1}^{n} \log \mathcal{N}(y_i \mid \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top, \beta^{-1})$$
$$= \frac{n}{2} \log \beta - \frac{n}{2} \log(2\pi) - \beta E_D(\mathbf{w}),$$

where $E_D(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top)^2$.

- Maximizing log-likelihood $\equiv$ minimizing $E_D(\mathbf{w})$:

$$\mathbf{w}_{ML} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{Y}_n, \; \boldsymbol{\Phi} = \{(\boldsymbol{\phi}_j(\mathbf{x}_i))_{j=0}^{M-1}\}_{i=1}^{n},$$
$$\frac{1}{\beta_{ML}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{w}_{ML} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^{\mathrm{T}})^2.$$

Least Squares = MLE

- Regularized Least Squares

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \to \min_{\mathbf{w}};$$

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}\cdot\phi(\mathbf{x}_i)^{\mathrm{T}})^2 + \frac{\lambda}{2}\mathbf{w}\cdot\mathbf{w}^{\mathrm{T}} \to \min_{\mathbf{w}}.$$

- Solution has the form

$$\mathbf{w}_{LS} = \left(\lambda\mathbf{I} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{Y}_n.$$

Parameter distribution

- We have a data sample $\mathcal{D}_n = (\mathbf{X}_n, \mathbf{Y}_n)$ from a linear basis function model.
- Likelihood:

$$p(\mathcal{D}_n \mid \mathbf{w}) = \prod_{i=1}^{n} \mathcal{N}(y_i \mid \mathbf{w} \cdot \phi(\mathbf{x}_i)^\top, \beta^{-1}).$$

- Thus the likelihood is Gaussian:

$$p(\mathcal{D}_n \mid \mathbf{w}) = \mathcal{N}(\mathbf{Y}_n \mid \mathbf{\Phi} \cdot \mathbf{w}^\top, \beta^{-1}\mathbf{I}).$$

- The typical prior is Gaussian as well:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I}).$$

Conditional Gaussian distribution

We assume that

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}),$$

$$p(\mathbf{y} \mid \mathbf{z}) = \mathcal{N}(\mathbf{y} \mid \mathbf{A}\mathbf{z} + \mathbf{b}, \mathbf{L}^{-1}).$$

Then we can prove that

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}),$$

$$p(\mathbf{z} \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{z} \mid \boldsymbol{\Sigma}\left[\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right], \boldsymbol{\Sigma}\right),$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}.$$

Parameter distribution

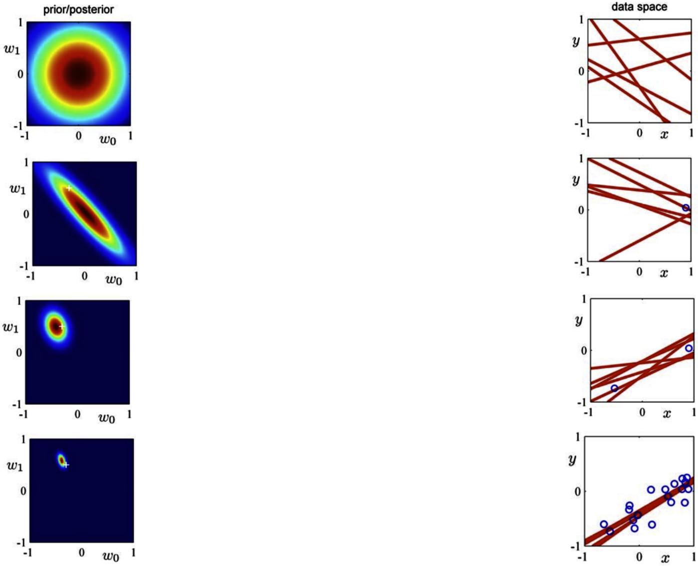
- Thus the posterior is defined by

$$
\begin{aligned}
p(\mathbf{w} \mid \mathcal{D}_n) &= \mathcal{N}(\mathbf{w} \mid \boldsymbol{\omega}_n, \mathbf{S}_n), \\
\mathbf{S}_n &= \left(\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1}, \\
\boldsymbol{\omega}_n &= \beta \mathbf{S}_n \boldsymbol{\Phi}^\top \mathbf{Y}_n.
\end{aligned}
$$

- The log posterior

$$
\log p(\mathbf{w} \mid \mathcal{D}_n) = -\frac{\beta}{2} \sum_{i=1}^{n} \{y_i - \mathbf{w}^\top \phi(\mathbf{x}_i)\}^2 - \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} + \text{const}.
$$

Sequential Bayesian Learning



Predictive Distribution

- Make prediction of y for new value of $\mathbf{x}$:

$$p(y \mid \mathbf{x}, \mathcal{D}_n, \alpha, \beta) = \int p(y \mid \mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w} \mid \mathcal{D}_n, \alpha, \beta) d\mathbf{w}.$$

- Actually, posterior of $\mathbf{w}$ is $p(\mathbf{w} \mid \mathcal{D}_n) = \mathcal{N}(\mathbf{w} \mid \boldsymbol{\omega}_n, \mathbf{S}_n)$ with
  - $\mathbf{S}_n = \left(\alpha^{-1}\mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1}$ — posterior covariance of $\mathbf{w}$;
  - $\boldsymbol{\omega}_n = \beta \mathbf{S}_n \boldsymbol{\Phi}^\top \mathbf{Y}_n$ — posterior mean of $\mathbf{w}$.
- Since $p(y \mid \mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y \mid f(\mathbf{x}, \mathbf{w}), \beta^{-1})$, then

$$p(y \mid \mathbf{x}, \mathcal{D}_n, \alpha, \beta) = \mathcal{N}(y \mid \boldsymbol{\omega}_n \cdot \boldsymbol{\phi}(\mathbf{x})^\top, \sigma_n^2(\mathbf{x})).$$

  Here

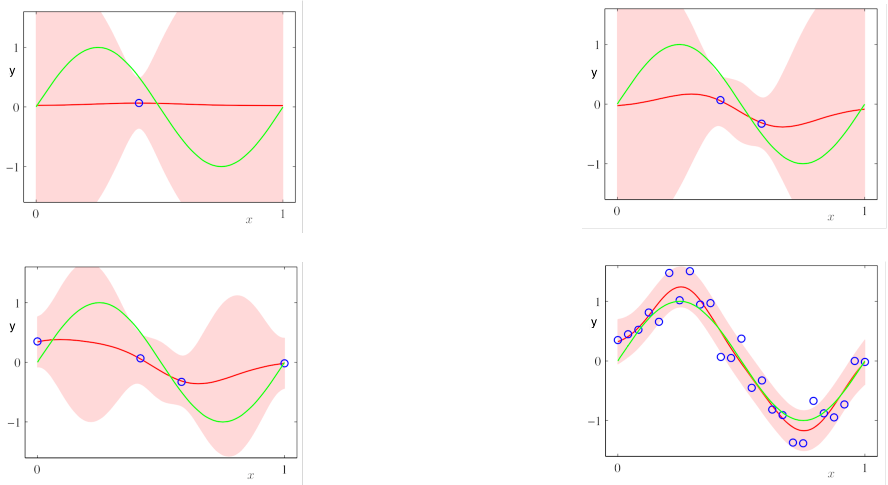
$$\sigma_n^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_n \boldsymbol{\phi}(\mathbf{x}).$$

- We can use posterior mean for point prediction

$$\widehat{f}(\mathbf{x}, \mathbf{w}) = \boldsymbol{\omega}_n \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

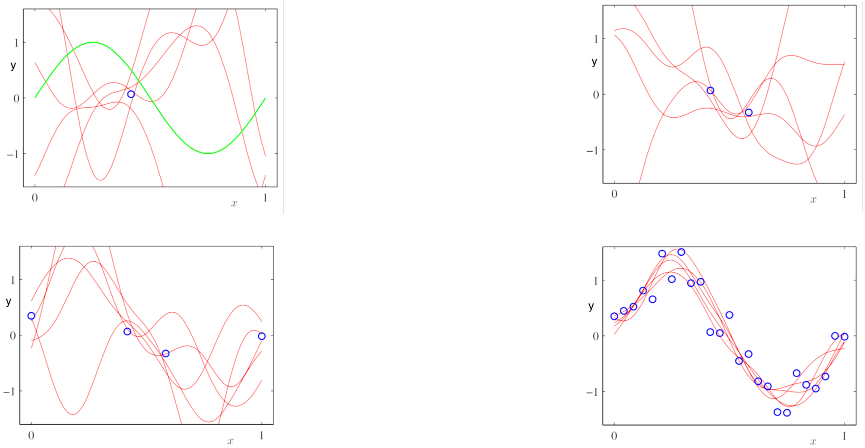
  and posterior variance $\sigma_n^2(\mathbf{x})$ for its uncertainty estimate.

Predictive Distribution



$M = 9$ Gaussian basis functions were used as $\phi(\mathbf{x})$

Samples from the Predictive Distribution



Plots of $f(\mathbf{x}, \mathbf{w})$ using samples from the posterior distributions over $\mathbf{w} \sim p(\mathbf{w} \mid \mathcal{D}_n, \alpha, \beta)$ for some α and β.

Predictive Distribution

- Make prediction of y for new value of $\mathbf{x}$:

$$p(y \mid \mathbf{x}, \mathcal{D}_n, \alpha, \beta) = \int p(y \mid \mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w} \mid \mathcal{D}_n, \alpha, \beta) d\mathbf{w}.$$

  Depends on α and β! How to define them? $\Rightarrow$ Full Bayesian approach!
- We introduce hyperpriors over α and β

$$p(y \mid \mathbf{x}, \mathcal{D}_n) = \int\int\int p(y \mid \mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w} \mid \mathcal{D}_n, \alpha, \beta) p(\alpha, \beta \mid \mathcal{D}_n) d\mathbf{w} d\alpha d\beta.$$

- We assume that the posterior distribution $p(\alpha, \beta \mid \mathcal{D}_n)$ is sharply peaked around values $\hat{\alpha}$ and $\hat{\beta}$.
- Then we simply marginalize over $\mathbf{w}$, where α and β are fixed to the values $\hat{\alpha}$ and $\hat{\beta}$, so that

$$p(y \mid \mathbf{x}, \mathcal{D}_n) \approx p(y \mid \mathbf{x}, \mathcal{D}_n, \hat{\alpha}, \hat{\beta}) = \int p(y \mid \mathbf{x}, \mathbf{w}, \hat{\beta}) p(\mathbf{w} \mid \mathcal{D}_n, \hat{\alpha}, \hat{\beta}) d\mathbf{w}.$$

Model Selection for Bayesian Regression

- The posterior for α and β is given by

$$p(\alpha, \beta \mid \mathcal{D}_n) \sim p(\mathcal{D}_n \mid \alpha, \beta) \cdot p(\alpha, \beta).$$

- If the prior $p(\alpha, \beta)$ is relatively flat, then approximately

$$(\hat{\alpha}, \hat{\beta}) = \arg\max_{\alpha, \beta} p(\mathcal{D}_n \mid \alpha, \beta).$$

- To obtain $(\hat{\alpha}, \hat{\beta})$ iterative optimization is used!

Approximate Bayesian Inference Problem

Posterior distribution:

$$p(\mathbf{w} \mid \mathcal{D}_n) = \frac{p(\mathcal{D}_n \mid \mathbf{w})\ p(\mathbf{w})}{p(\mathcal{D}_n)},$$

The problem with exact posterior computation comes from the denominator:

$$p(\mathcal{D}_n) = \int p(\mathcal{D}_n, \mathbf{w}) d\mathbf{w}$$

as

- generally, the complexity of this integral computation grows exponentially with dimensionality;
- an exception is the case of conjugate pairs of prior and likelihood.

On the next lecture we will discuss approximate Bayesian inference:

- MCMC (Markov Chain Monte Carlo);
- Variational Inference.

Thank you for your attention![1]

[1] Slides are partially based on the material provided by Evgeny Burnaev.