Lecture 16-17

CM50264: Machine Learning 1
Bayesian Linear Regression and Gaussian Process Regression

Kwang In Kim

Previously in machine learning...
Regularized linear least-squares regression

Data:

$$D = \{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^N, y^N)\} \subset \mathbb{R}^n \times \mathbb{R}.$$

Linear regression function:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}.$$

- (Plain) linear least-squares regression minimizes

$$\mathcal{O}(\mathbf{w}) = \sum_{i=1}^{N} (\mathbf{w}^\top \mathbf{x}^i - y^i)^2$$

- Regularized linear least-squares regression minimizes

$$\mathcal{O}(\mathbf{w}) = \sum_{i=1}^{N} (\mathbf{w}^\top \mathbf{x}^i - y^i)^2 + \lambda \|\mathbf{w}\|^2,$$

for the regularization (hyper-)parameter $\lambda \geq 0$.

Regularized linear least-squares regression

Data matrix: $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^N]$; Label vector: $\mathbf{y} = [y^1, \ldots, y^N]^\top$.

- (Plain) linear least-squares regression minimizes

$$\mathcal{O}(\mathbf{w}) = \sum_{i=1}^{N} (\mathbf{w}^\top \mathbf{x}^i - y^i)^2 = \|\mathbf{X}^\top \mathbf{w} - \mathbf{y}\|^2.$$

The minimizer $\mathbf{w}^*$ is obtained by solving a linear system

$$\mathbf{X}\mathbf{X}^\top \mathbf{w} = \mathbf{X}\mathbf{y}.$$

For high-dimensional problems ($N < n$), $\mathbf{X}\mathbf{X}^\top$ is rank deficient: Infinitely many solutions exist.

- Regularized linear least-squares regression minimizes for $\lambda \geq 0$

$$\mathcal{O}(\mathbf{w}) = \sum_{i=1}^{N} (\mathbf{w}^\top \mathbf{x}^i - y^i)^2 + \lambda \|\mathbf{w}\|^2 = \|\mathbf{X}^\top \mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|^2.$$

The minimizer $\mathbf{w}^*$ is obtained by solving a linear system

$$(\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I})\mathbf{w} = \mathbf{X}\mathbf{y}.$$

For $\lambda > 0$, $\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}$ is always full rank: A unique solution exists.

(Deterministic) linear regression summary

- Input: Data $\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^N, y^N)\} \subset \mathbb{R}^n \times \mathbb{R}$; regularization parameter $\lambda \geq 0$.
- Training:
  - Build the data matrix: $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^N]$ and label vector $\mathbf{y} = [y^1, \ldots, y^N]$;
  - Solve a linear system to obtain $\mathbf{w}^*$: $(\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I})\mathbf{w} = \mathbf{X}\mathbf{y}$;
- Testing: $f(\mathbf{x}') = (\mathbf{w}^*)^\top \mathbf{x}'$.

Probabilistic setup revisited

We will discuss probabilistic interpretations of (plain) linear regression and regularized linear regression algorithms (see ‘L09 Regularisation & Model Types’ slides).

- Input and output variables $\mathbf{x}$ and y are random variables.
- Training data $\mathcal{D} = \{(\mathbf{x}^N, y^N), \ldots, (\mathbf{x}^N, y^N)\}$ is sampled from an unknown probability distribution $p(\mathbf{x}, y)$.
- There is an underlying ground-truth function $f^*(\mathbf{x}) = y$, but our observations (training data) are noisy:

$$y^i = f^*(\mathbf{x}^i) + \text{noise}^i, 1 \leq i \leq N.$$

- noise^i is a random variable.

Independent and identically distributed (i.i.d.) Gaussian noise model

$$y^i = f(\mathbf{x}^i) + \text{noise}^i$$

- noise^i represents the deviation between the observed label y^i and the prediction $f(\mathbf{x}^i)$: training error.
- noise^i is independent of noise^j ($i \neq j$).
- Distribution of noise^i is Gaussian $\mathcal{N}(\mu, \sigma^2)$ with
  - mean μ zero
  - variance σ^2 identical across i
- In linear regression:

$$\begin{aligned} y^i &= f(\mathbf{x}^i) + \text{noise} \\ &= \mathbf{w}^\top \mathbf{x}^i + \text{noise}. \end{aligned}$$

Why i.i.d. Gaussian noise model?

Maximum likelihood (ML) estimation

Training data: $\mathcal{D} = \{(\mathbf{x}^N, y^N), \ldots, (\mathbf{x}^N, y^N)\}$.
Data matrix: $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^N]$.
Label vector: $\mathbf{y} = [y^1, \ldots, y^N]^\top$.
Our model:

$$y^i = \mathbf{w}^\top \mathbf{x}^i + \text{noise}$$
$$\text{noise} \sim \mathcal{N}(0, \sigma^2).$$

The maximum likelihood (ML) estimation chooses $\mathbf{w}^*$ that maximizes the likelihood of $\mathbf{w}$ given $\mathcal{D}$, the possibility of observing training data points $\mathcal{D}$ given the hypothesized solution $\mathbf{w}$:

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \mathbf{w}) &= \prod_{i=1}^{N} p(y^i|\mathbf{x}^i, \mathbf{w}) \text{ (i.i.d. noise)} \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{w}^\top \mathbf{x}^i - y^i)^2}{2\sigma^2}\right) \text{ (Gaussian noise)} \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \exp\left(-\frac{\|\mathbf{X}^\top \mathbf{w} - \mathbf{y}\|^2}{2\sigma^2}\right). \end{aligned}$$

Maximum likelihood (ML) estimation

Our likelihood model:

$$p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \exp\left(-\frac{\|\mathbf{X}^\top\mathbf{w}-\mathbf{y}\|^2}{2\sigma^2}\right).$$

ML estimation maximizes $p(\mathbf{y}|\mathbf{X},\mathbf{w})$:

$$\begin{aligned}
\mathbf{w}^* &= \operatorname*{arg\,max}_{\mathbf{w}\in\mathbb{R}^n} \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \exp\left(-\frac{\|\mathbf{X}^\top\mathbf{w}-\mathbf{y}\|^2}{2\sigma^2}\right)\\
&= \operatorname*{arg\,max}_{\mathbf{w}\in\mathbb{R}^n} \exp\left(-\frac{\|\mathbf{X}^\top\mathbf{w}-\mathbf{y}\|^2}{2\sigma^2}\right)\\
&= \operatorname*{arg\,max}_{\mathbf{w}\in\mathbb{R}^n} \left(-\frac{\|\mathbf{X}^\top\mathbf{w}-\mathbf{y}\|^2}{2\sigma^2}\right)\\
&= \operatorname*{arg\,min}_{\mathbf{w}\in\mathbb{R}^n} \frac{\|\mathbf{X}^\top\mathbf{w}-\mathbf{y}\|^2}{2\sigma^2}\\
&= \operatorname*{arg\,min}_{\mathbf{w}\in\mathbb{R}^n} \|\mathbf{X}^\top\mathbf{w}-\mathbf{y}\|^2
\end{aligned}$$

$$\Leftrightarrow \mathbf{X}\mathbf{X}^\top\mathbf{w}^* = \mathbf{X}\mathbf{y}.$$

ML under i.i.d. Gaussian noise is the same as least-squares regression.

Maximum a posteriori (MAP) estimation

In ML, we maximize the likelihood:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}).$$

In MAP, we maximize the posterior:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{X}, \mathbf{y}).$$

Maximum a posteriori (MAP) estimation

Applying Bayes' rule,

$$\begin{aligned}\arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{y},\mathbf{X}) &= \arg\max_{\mathbf{w}} \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w}|\mathbf{X})}{p(\mathbf{y}|\mathbf{X})} \\ &= \arg\max_{\mathbf{w}} \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})} \\ &= \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w}).\end{aligned}$$

We know how to calculate $p(\mathbf{y}|\mathbf{X},\mathbf{w})$. What about $p(\mathbf{w})$?

This is where we apply our a priori knowledge of **w**.

What priori knowledge?

Gaussian prior

$$p(\mathbf{w}|\mathbf{X},\mathbf{y}) \propto p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})$$

If we assume that $p(\mathbf{w})$ is a centered Gaussian $\mathcal{N}(0,\mathbf{I})$:

$$p(\mathbf{w}) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{\|\mathbf{w}\|^2}{2}\right),$$

maximizing the posterior $p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})$ biases the solution $\mathbf{w}^*$ towards 0:

$$p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w}) = \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \exp\left(-\frac{\|\mathbf{X}^\top\mathbf{w}-\mathbf{y}\|^2}{2\sigma^2}\right) \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{\|\mathbf{w}\|^2}{2}\right).$$

$$\begin{aligned}
\arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w}) &= \arg\max_{\mathbf{w}} \exp\left(-\frac{\|\mathbf{X}^\top\mathbf{w}-\mathbf{y}\|^2}{2\sigma^2}\right) \exp\left(-\frac{\|\mathbf{w}\|^2}{2}\right) \\
&= \arg\max_{\mathbf{w}} \exp\left(-\frac{\|\mathbf{X}^\top\mathbf{w}-\mathbf{y}\|^2}{2\sigma^2} - \frac{\|\mathbf{w}\|^2}{2}\right) \\
&= \arg\max_{\mathbf{w}} \left(-\frac{\|\mathbf{X}^\top\mathbf{w}-\mathbf{y}\|^2}{2\sigma^2} - \frac{\|\mathbf{w}\|^2}{2}\right) \\
&= \arg\min_{\mathbf{w}} \frac{\|\mathbf{X}^\top\mathbf{w}-\mathbf{y}\|^2}{2\sigma^2} + \frac{\|\mathbf{w}\|^2}{2} \\
&= \arg\min_{\mathbf{w}} \|\mathbf{X}^\top\mathbf{w}-\mathbf{y}\|^2 + \sigma^2\|\mathbf{w}\|^2.
\end{aligned}$$

Maximum a posteriori estimation with Gaussian prior

$$p(\mathbf{w}|\mathbf{X},\mathbf{y}) \propto p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})$$

Assuming that $p(\mathbf{w})$ is a Gaussian $\mathcal{N}(0,\mathbf{I})$, maximizing $p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})$ biases the solution $\mathbf{w}^*$ towards 0.

$$\begin{aligned}\mathbf{w}^* &= \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w}) \\ &= \arg\min_{\mathbf{w}} \|\mathbf{X}^\top\mathbf{w} - \mathbf{y}\|^2 + \sigma^2\|\mathbf{w}\|^2 \\ &\vdots\end{aligned}$$

$$\Leftrightarrow (\mathbf{X}\mathbf{X}^\top + \sigma^2 I)\mathbf{w}^* = \mathbf{X}\mathbf{y}.$$

With a Gaussian prior and i.i.d. Gaussian noise model, MAP estimate becomes regularized least-squares solution with σ^2 as the regularization hyper-parameter.

Why Gaussian prior?

MAP linear regression summary

- Input: Data $\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^N, y^N)\} \subset \mathbb{R}^n \times \mathbb{R}$; noise parameter $\sigma^2 \geq 0$.
- Training:
    - Build the data matrix: $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^N]$ and label vector $\mathbf{y} = [y^1, \ldots, y^N]$;
    - Solve a linear system to obtain $\mathbf{w}^*$: $(\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I})\mathbf{w} = \mathbf{X}\mathbf{y}$;
- Testing: $f(\mathbf{x}') = (\mathbf{w}^*)^\top \mathbf{x}'$.

MAP linear regression summary

We choose $\mathbf{w}^*$ by maximizing $p(\mathbf{w}|\mathbf{y}, \mathbf{X})$

and then,

apply the resulting linear regression function f to a new data point $\mathbf{x}'$: $f(\mathbf{x}') = (\mathbf{w}^*)^\top \mathbf{x}'$.

Bayesian linear regression: basic idea

We don't really need to build f (equivalently $\mathbf{w}^*$) explicitly as an intermediate result

if we just want to make a prediction y' for a given input $\mathbf{x}'$ (or inputs):
We can maximize the posterior (or predictive distribution):[1]

$$p(y'|\mathbf{x}', \mathbf{y}, \mathbf{X}).$$

Rewrite this in plain text!

[1] Cf. the parameter posterior $p(\mathbf{w}|\mathbf{y}, \mathbf{X})$.

Marginalization

From a given joint distribution $p(\mathbf{a}, \mathbf{b})$,
one can calculate the marginal distribution $p(\mathbf{a})$ by integrating **b** out:

$$p(\mathbf{a}) = \int p(\mathbf{a}, \mathbf{b}) d\mathbf{b}.$$

Similarly,

$$\begin{aligned} p(\mathbf{a}, \mathbf{c}) &= \int p(\mathbf{a}, \mathbf{b}, \mathbf{c}) d\mathbf{b} \\ &= \int p(\mathbf{a}|\mathbf{b}, \mathbf{c}) p(\mathbf{b}|\mathbf{c}) d\mathbf{b}. \end{aligned}$$

Bayesian linear regression

Applying the marginalization of $\mathbf{w}$ to the predictive distribution

$$p(y'|\mathbf{x}',\mathbf{y},\mathbf{X}) = \int p(y'|\mathbf{x}',\mathbf{w})p(\mathbf{w}|\mathbf{y},\mathbf{X})d\mathbf{w}$$

and combining it with the parameter posterior and likelihood

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) = \mathcal{N}\left((\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{X}\mathbf{y}, \sigma^2(\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I})^{-1}\right) \qquad (1)$$

$$p(y'|\mathbf{x}',\mathbf{w}) = \mathcal{N}(\mathbf{w}^\top\mathbf{x}', \sigma^2),$$

we obtain

$$p(y'|\mathbf{x}',\mathbf{y},\mathbf{X}) = \mathcal{N}\left(\mathbf{x}'^\top(\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{X}\mathbf{y}, \mathbf{x}'^\top\sigma^2(\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{x}'\right). \qquad (2)$$

Prove Eqs. 1 and 2.

Predictive distribution

$$p(y|\mathbf{x}', \mathbf{y}, \mathbf{X}) = \mathcal{N}\left(\mathbf{x}'^{\top}(\mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I})^{-1}\mathbf{X}\mathbf{y}, \mathbf{x}'^{\top}\sigma^2(\mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I})^{-1}\mathbf{x}'\right) \tag{3}$$

$$p(y'|\mathbf{x}', \mathbf{y}, \mathbf{X}) = \int p(y'|\mathbf{x}', \mathbf{w})p(\mathbf{w}|\mathbf{y}, \mathbf{X})d\mathbf{w}. \tag{4}$$

- The output of Bayesian linear regression is a probability distribution: Gaussian for Gaussian prior + Gaussian noise.
- If we take the mean of this predictive distribution, the result is the same as the MAP solution:

$$\mathbf{x}'^{\top}(\mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I})^{-1}\mathbf{X}\mathbf{y} = \mathbf{x}'^{\top}\mathbf{w}^*,$$

$$\mathbf{w}^* = (\mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I})^{-1}\mathbf{X}\mathbf{y}.$$

- Equation 3 represents the posterior $p(y|\mathbf{x}', \mathbf{y}, \mathbf{X})$ without explicitly involving the parameter vector $\mathbf{w}^*$. This does not mean that our regression model

$$f(\mathbf{x}) = \mathbf{w}^{\top}\mathbf{x}$$

is removed: The posterior should be consistent with the marginalization rule (Eq. 4).

Mean = mode for Gaussian.

Predictive distribution

$$p(y|\mathbf{x}', \mathbf{y}, \mathbf{X}) = \mathcal{N}\left(\mathbf{x}'^{\top}(\mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I})^{-1}\mathbf{X}\mathbf{y}, \mathbf{x}'^{\top}\sigma^2(\mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I})^{-1}\mathbf{x}'\right).$$

The prediction is a Gaussian distribution characterized by

predictive mean: $\mathbf{x}'^{\top}(\mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I})^{-1}\mathbf{X}\mathbf{y}$

predictive variance: $\mathbf{x}'^{\top}\sigma^2(\mathbf{X}\mathbf{X}^{\top} + \sigma^2\mathbf{I})^{-1}\mathbf{x}'$.

The predictive variance represents how confident the prediction is

- Large variance $\rightarrow$ low confidence.
- Under the i.i.d. Gaussian noise model, predictive variances are independent of training labels **y** (and underlying *f*). For other noise models, predictive variances might depend on **y**.
- Similarly to the mean prediction, our predictive variance is limited by the model assumption:
  If $\mathbf{x}' = 0$, we have an absolutely confident prediction. Why is 0 special?

Bayesian linear regression

Demo

Marginal likelihood

How can we choose the noise hyperparameter σ^2 (or equivalently λ)?

When we were applying Bayes' rule

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})},$$

we discarded $p(\mathbf{y}|\mathbf{X})$ since it is independent of **w**.

The marginal likelihood $p(\mathbf{y}|\mathbf{X})$ is a function of σ^2:

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})d\mathbf{w}.$$

This represents how well $p(\mathbf{w})$ respects the observed data and it can be used as a criteria for optimizing σ^2.

Bayesian linear regression summary

- Input: Data $\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^N, y^N)\} \subset \mathbb{R}^n \times \mathbb{R}$; noise parameter $\sigma^2 \geq 0$.
- Construct the predictive distribution $p(y'|\mathbf{x}', \mathbf{y}, \mathbf{X})$ for a given input $\mathbf{x}'$:

$$p(y|\mathbf{x}', \mathbf{y}, X) = \mathcal{N}\left(\mathbf{x}'^\top \mathbf{A}\mathbf{X}\mathbf{y}), \mathbf{x}'^\top \sigma^2 \mathbf{A}\mathbf{x}'\right)$$
$$\mathbf{A} = (\mathbf{X}\mathbf{X}^\top + \sigma^2 \mathbf{I})^{-1}.$$

- No clear distinction of training and testing stages; **AXy** could be pre-calculated.

Bayesian non-linear regression

Idea: map $\mathbf{x}$ to a feature space $\mathcal{F}$ using a non-linear map ϕ and build a linear regressor in $\mathcal{F}$:

$$f(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}).$$

$$\begin{aligned} p(y'|\mathbf{x}', \mathbf{y}, \mathbf{X}) &:= p(y'|\mathbf{x}', \mathbf{y}, \mathbf{\Phi}) \\ &= \mathcal{N}\Big(\phi(\mathbf{x}')^\top(\mathbf{\Phi}\mathbf{\Phi}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{\Phi}\mathbf{y}, \\ &\qquad\qquad \phi(\mathbf{x}')^\top(\mathbf{\Phi}\mathbf{\Phi}^\top + \sigma^2\mathbf{I})^{-1}\sigma^2\phi(\mathbf{x}')\Big), \\ \mathbf{\Phi} &= [\phi(\mathbf{x}^1), \ldots, \phi(\mathbf{x}^N)]. \end{aligned}$$

Kernelization

$$
\begin{aligned}
& p(y'|\mathbf{x}', \mathbf{y}, \Phi) \\
& = \mathcal{N}\Big(\phi(\mathbf{x}')^\top (\Phi\Phi^\top + \sigma^2\mathbf{I})^{-1}\Phi\mathbf{y}, \phi(\mathbf{x}')^\top \sigma^2 (\Phi\Phi^\top + \mathbf{I})^{-1}\phi(\mathbf{x}')\Big), \\
& = \mathcal{N}\Big(\phi(\mathbf{x}')^\top \Phi(\Phi^\top\Phi + \sigma^2\mathbf{I})^{-1}\mathbf{y}, \\
& \qquad \phi(\mathbf{x}')^\top\phi(\mathbf{x}') - \phi(\mathbf{x}')^\top\Phi(\Phi^\top\Phi + \sigma^2\mathbf{I})^{-1}\Phi^\top\phi(\mathbf{x}')\Big), \\
& \Phi = [\phi(\mathbf{x}^1), \ldots, \phi(\mathbf{x}^N)].
\end{aligned}
$$

ϕ is always given in the inner-product form $\phi(\mathbf{a})^\top\phi(\mathbf{b})$.

see Sherman-Morrison-Woodbury formula (last slide) for the second equality.

Kernelization

Using a positive definite kernel function $k(\mathbf{a}, \mathbf{b}) = \phi(\mathbf{a})^\top \phi(\mathbf{b})$, we obtain

$$\begin{aligned} p(y'|\mathbf{x}', \mathbf{y}, \Phi) = \mathcal{N}\Big(& \phi(\mathbf{x}')^\top \Phi(\Phi^\top \Phi + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, \\ & \phi(\mathbf{x}')^\top \phi(\mathbf{x}') - \phi(\mathbf{x}')^\top \Phi(\Phi^\top \Phi + \sigma^2 \mathbf{I})^{-1} \Phi^\top \phi(\mathbf{x}')\Big), \\ = \mathcal{N}\Big(& \mathbf{k}^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, k(\mathbf{x}', \mathbf{x}') - \mathbf{k}^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}\Big), \end{aligned}$$

$$\begin{aligned} \mathbf{k} &= [k(\mathbf{x}', \mathbf{x}^1), \ldots, k(\mathbf{x}', \mathbf{x}^N)]^\top \\ [\mathbf{K}]_{i,j} &= k(\mathbf{x}^i, \mathbf{x}^j). \end{aligned}$$

Two modes of Bayesian non-linear regression

Original: complexity $\mathcal{O}(n^3)$ ($n = \dim(\mathcal{F})$)

$$p(y|\mathbf{x}', \mathbf{y}, \mathbf{X}) := p(y|\mathbf{x}', \mathbf{y}, \mathbf{\Phi})$$
$$= \mathcal{N}\Big(\sigma^{-2}\phi(\mathbf{x}')^\top(\mathbf{\Phi}\mathbf{\Phi}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{\Phi}\mathbf{y}, \phi(\mathbf{x}')^\top\sigma^2(\mathbf{\Phi}\mathbf{\Phi}^\top + \sigma^2\mathbf{I})^{-1}\phi(\mathbf{x})\Big).$$

Kernelized version: complexity $\mathcal{O}(N^3)$ ($N = \#$ data points)

$$p(y|\mathbf{x}', \mathbf{y}, \mathbf{X}) = \mathcal{N}\Big(\mathbf{k}^\top(K + \sigma^2\mathbf{I})^{-1}\mathbf{y}, k(\mathbf{x}', \mathbf{x}') - \mathbf{k}^\top(K + \sigma^2\mathbf{I})^{-1}\mathbf{k}\Big).$$

When $n > N$, kernelized version is preferable.

Gaussian kernels

$$k(\mathbf{a}, \mathbf{b}) = \exp\left(-\frac{\|\mathbf{a} - \mathbf{b}\|^2}{\sigma_k^2}\right)$$

- Simple linear regression is powerful in high-dimensional spaces (see ‘L15 Regularized Regression and Support Vector Machines’ slides).
- For any positive definite kernel k, there is a (non-unique) feature map $\phi : \mathbb{R}^n \to \mathcal{H}$ such that $k(\mathbf{a}, \mathbf{b}) = \phi(\mathbf{a})^\top \phi(\mathbf{b})$.
- For the Gaussian kernel, the dimensionality of the feature space $\mathcal{H}$ is infinite.

Bayesian nonlinear regression summary

- Input: Data $\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^N, y^N)\} \subset \mathbb{R}^n \times \mathbb{R}$; noise parameter $\sigma^2 \geq 0$.
- Construct the predictive distribution $p(y'|\mathbf{x}', \mathbf{y}, \mathbf{X})$ for a given input $\mathbf{x}'$:

$$
\begin{aligned}
p(y'|\mathbf{x}', \mathbf{y}, \mathbf{X}) =& \mathcal{N}\Big(\mathbf{k}^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, \\
& \qquad k(\mathbf{x}', \mathbf{x}') - \mathbf{k}^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}\Big).
\end{aligned}
$$

- No clear distinction of training and testing stages; $(\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}$ could be pre-calculated.

The deterministic part (i.e. predictive mean) of kernelized Bayesian nonlinear regression is also called kernel ridge regression and regularization networks.

Demo

- Effect of varying hyper-parameters, noise variance σ^2 and kernel parameter σ_k^2?
- How do we select hyper-parameters?
  The marginal likelihood $p(\mathbf{y}|\mathbf{X})$ is a function of σ^2 and σ_k^2:

  $$p(\mathbf{y}|\mathbf{X}) := p(\mathbf{y}|\boldsymbol{\Phi}) = \int p(\mathbf{y}|\boldsymbol{\Phi}, \mathbf{w})p(\mathbf{w})d\mathbf{w}.$$

  This represents how well $p(\mathbf{w})$ respects the observed data and it can be used as a criteria for optimizing σ^2 and σ_k^2 (see 'Calculating the marginal likelihood' in the last slide).

This is not a truly Bayesian approach. How do we choose the hyper-parameters in a fully Bayesian way?

Parametric regression

For Bayesian linear regression, we adopted the model assumption

$$f(\mathbf{x}) = (\mathbf{w}^*)^\top \mathbf{x}$$

and made predictions using the marginalization

$$p(y'|\mathbf{x}', \mathbf{y}, \mathbf{X}) = \int p(y'|\mathbf{x}', \mathbf{w}) p(\mathbf{w}|\mathbf{y}, \mathbf{X}) d\mathbf{w}.$$

The parameter posterior $p(\mathbf{w}|\mathbf{y}, \mathbf{X})$ can be evaluated by combining the likelihood $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$ and the prior $p(\mathbf{w}) = p(\mathbf{w}|\mathbf{X})$.

We used a centered Gaussian $\mathcal{N}(0, \mathbf{I})$ prior:

$$p(\mathbf{w}) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{\|\mathbf{w}\|^2}{2}\right).$$

Non-parametric regression

If we remove the model assumption $f(\mathbf{x}) = (\mathbf{w}^*)^\top \mathbf{x}$ and use f as a variable,
our prediction rule will look like

$$p(y'|\mathbf{x}', \mathbf{y}, \mathbf{X}) = \int p(y'|\mathbf{x}', f)p(f|\mathbf{y}, \mathbf{X})df.$$

The function posterior $p(f|\mathbf{y}, \mathbf{X})$ depends on the likelihood $p(\mathbf{y}|f, \mathbf{X})$ and the prior $p(f|\mathbf{X})$.

Now we need a Gaussian distribution $p(f|\mathbf{X})$ on the space of functions.

Gaussian random vectors

- A Gaussian random variable w follows a Gaussian distribution:

$$w \sim \mathcal{N}(\mu, \sigma^2)$$

$$p(w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(w-\mu)^2}{2\sigma^2}\right)$$

  with mean μ and variance σ^2.

- A Gaussian random vector $\mathbf{w} \in \mathbb{R}^n$ is a collection of random variables $\{\mathbf{w}_j\}_{j=1}^n$ that has a joint Gaussian distribution

$$\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

$$p(\mathbf{w}) = \frac{1}{\sqrt{(2\pi\sigma^2)^n|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{w}-\boldsymbol{\mu})\right)$$

  with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.
  Elements $\{\mathbf{w}_j\}_{j=1}^n$ of $\mathbf{w}$ is indexed by an integer $j \in 1, \ldots, n$.

Gaussian processes

A Gaussian process (GP) f is a collection of random variables, any finite subset of which has a joint Gaussian distribution.

- A GP is specified by a mean function $m(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$:

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]. \end{aligned}$$

- The elements of a GP f is indexed by a continuous variable $\mathbf{x} \in \mathbb{R}^n$.
- For any set $\mathbf{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$, $\mathbf{f}_\mathbf{X} =: \{f(\mathbf{x}^1), \ldots, f(\mathbf{x}^N)\}$ is a Gaussian random vector characterized by
  mean vector $\boldsymbol{\mu}_\mathbf{X} = [m(\mathbf{x}^1), \ldots, m(\mathbf{x}^N)]^\top$
  covariance matrix $\mathbf{K}_\mathbf{X} : [\mathbf{K}_\mathbf{X}]_{i,j} = k(\mathbf{x}^i, \mathbf{x}^j)$.

Gaussian processes

- A GP is specified by a mean function $m(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$:

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]. \end{aligned}$$

- For any set $\mathbf{X} = \{\mathbf{x}^1, \dots, \mathbf{x}^N\}$,
  $\mathbf{f}_\mathbf{X} =: \{f(\mathbf{x}^1), \dots, f(\mathbf{x}^N)\}$ is a Gaussian random vector characterized by
  mean vector $\boldsymbol{\mu}_\mathbf{X} = [m(\mathbf{x}^1), \dots, m(\mathbf{x}^N)]^\top$
  covariance matrix $\mathbf{K}_\mathbf{X} : [\mathbf{K}_\mathbf{X}]_{i,j} = k(\mathbf{x}^i, \mathbf{x}^j)$.
- A GP is the generalization of a Gaussian random vector to infinite-dimensional objects, e.g. functions:

$$f \sim \mathcal{GP}(m, k).$$

Gaussian processes regression

For $\overline{\mathbf{X}} := \mathbb{R}^n \backslash \{\mathbf{X}, \mathbf{x}'\}$ (all inputs other than training and test inputs),

$$\begin{aligned} p(y'|\mathbf{x}', \mathbf{y}, \mathbf{X}) &= \int p(y'|\mathbf{x}', f) p(f|\mathbf{y}, \mathbf{X}) df \\ &= \int p(y'|\mathbf{x}', \mathbf{f}_{\mathbf{X}}, \mathbf{f}_{\overline{\mathbf{X}}}) p(\mathbf{f}_{\mathbf{X}}, \mathbf{f}_{\overline{\mathbf{X}}}|\mathbf{y}, \mathbf{X}) d\mathbf{f}_{\mathbf{X}} d\mathbf{f}_{\overline{\mathbf{X}}} \\ &= \int p(y'|\mathbf{x}', \mathbf{f}_{\mathbf{X}}) p(\mathbf{f}_{\mathbf{X}}|\mathbf{y}, \mathbf{X}) d\mathbf{f}_{\mathbf{X}}. \end{aligned}$$

The third equality is called the marginalization property of GPs, generalizing

$$p(\mathbf{a}) = \int p(\mathbf{a}, \mathbf{b}) d\mathbf{b}.$$

Gaussian process regression

$$p(y'|\mathbf{x}', \mathbf{y}, \mathbf{X}) = \int p(y'|\mathbf{x}', \mathbf{f_x}) p(\mathbf{f_x}|\mathbf{y}, \mathbf{X}) d\mathbf{f_x}$$

$$p(\mathbf{f_X}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{f_x}) p(\mathbf{f_x}|\mathbf{X})}{p(\mathbf{y}|\mathbf{X})}$$

- We place a zero mean GP prior on f:
  For any set $\mathbf{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$,
  $\mathbf{f_X}|\mathbf{X}$ is a Gaussian random vector:

$$p(\mathbf{f_x}|\mathbf{X}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_X).$$

- We place a zero-mean i.i.d. Gaussian noise model on $\mathbf{y}$:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{f_X}) = \mathcal{N}(\mathbf{f_X}, \sigma^2 \mathbf{I}).$$

Gaussian process regression

With the zero mean GP prior:

$$p(\mathbf{f_x}|\mathbf{X}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_X).$$

and i.i.d. Gaussian noise model:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{f_X}) = \mathcal{N}(\mathbf{f_X}, \sigma^2\mathbf{I}),$$

the joint distribution of $\mathbf{y}$ and $f(\mathbf{x}')$ is obtained as

$$\left(\begin{array}{c} \mathbf{y} \\ f(\mathbf{x}') \end{array} \right) \sim \mathcal{N}\left(\mathbf{0}, \left(\begin{array}{cc} (\mathbf{K}_X + \sigma^2\mathbf{I}) & \mathbf{k} \\ \mathbf{k}^\top & k(\mathbf{x}', \mathbf{x}') \end{array} \right)\right) \tag{5}$$

with $\mathbf{k} = [k(\mathbf{x}', \mathbf{x}^1), \ldots, k(\mathbf{x}', \mathbf{x}^N)]^\top$.

Using the Gaussian conditioning formula (last slide), we obtain the posterior

$$p(y'|\mathbf{x}', \mathbf{y}, \mathbf{X}) = \mathcal{N}\left(\mathbf{k}^\top(\mathbf{K_X} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, k(\mathbf{x}', \mathbf{x}') - \mathbf{k}^\top(\mathbf{K_X} + \sigma^2\mathbf{I})^{-1}\mathbf{k}\right).$$

How do we derive the joint distribution (Eq. 5)?

Equivalence of GP regression and kernelized Bayesian nonlinear regression

Loéve's theorem [Ber]:

$$\begin{aligned} & k(\cdot,\cdot) \text{ is a covariance function of a GP} \\ \Leftrightarrow\ & k(\cdot,\cdot) \text{ is a symmetric positive definite function (kernel).} \end{aligned}$$

Nonlinear feature map + linear Bayesian regression is the same as GP regression.

- Sherman-Morrison-Woodbury formula:
  For matrices $\mathbf{A} \in \mathbb{R}^{m \times m}, \mathbf{U} \in \mathbb{R}^{m \times n}, \mathbf{C} \in \mathbb{R}^{n \times n}, \mathbf{V} \in \mathbb{R}^{n \times m}$,

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U}\right)^{-1}\mathbf{VA}^{-1}$$

$$(\mathbf{A} + \mathbf{UCV})^{-1}\mathbf{UC} = \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U}\right)^{-1}.$$

  When $n < m$, $\left(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U}\right)^{-1}$ is less costly to calculate than $(\mathbf{A} + \mathbf{UCV})^{-1}$.

- Conditioning of a joint Gaussian is a Gaussian:

$$p\left(\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}\right) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{A} & \mathbf{C}^\top \\ \mathbf{C} & \mathbf{B} \end{bmatrix}\right)$$

$$\Rightarrow p(\mathbf{a}|\mathbf{b}) = \mathcal{N}(\mathbf{C}^\top\mathbf{B}^{-1}\mathbf{b}, \mathbf{A} - \mathbf{C}^\top\mathbf{B}^{-1}\mathbf{C}).$$

- Calculating the marginal likelihood:

$$-2\log p(\mathbf{y}|\mathbf{X}) = \mathbf{y}^\top(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y} + \log|\mathbf{K} + \sigma^2\mathbf{I}| + N\log 2\pi.$$

References

Ber Berlinet and Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability And Statistics*, Kluwer Academic, 2004.

Pet Petersen and Pedersen, *The Matrix Cookbook*

Teu Teukolsky, Vetterling, Flannery, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press (any edition)
`http://www.nr.com/`