Gaussian Processes

An Introduction to Probabilistic Machine Learning

Pablo Martínez Olmos

Machine Learning, Nov. 2019
Master in Information Health Engineering

Outline

Section 1

Probabilistic Machine Learning and Gaussian Models

Probabilistic modelling

- Probabilistic modelling emerged as one of the principal theoretical and practical approaches for designing machines that learn from data acquired through experience.
- Represent and manipulate uncertainty about models and predictions.
- Plays a central role in scientific data analysis, machine learning, robotics, cognitive science, and artificial intelligence.

Probabilistic modelling

- Learning can be thought of as inferring plausible models to explain observed data.
- Observed data can be consistent with many models, and therefore which model is appropriate given the data is uncertain.
- Similarly, predictions, about future data and the future consequences of actions, are uncertain.
- Probability theory provides a framework for modelling uncertainty.

REVIEW

doi:10.1038/nature14541

Probabilistic machine learning and artificial intelligence

Zoubin Ghahramani[1]

Gaussian Density

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- The sum of independent Gaussians is also Gaussian

$$p\left(\sum_{i=1}^{N} x_i\right) = \mathcal{N}\left(\sum_{i=1}^{N} x_i \,\middle|\, \sum_{i=1}^{N} \mu_i, \sum_{i=1}^{N} \sigma_i^2\right)$$

- Scaling a Gaussian leads to a Gaussian

$$p(cx) = \mathcal{N}(cx|c\mu, c^2\sigma^2)$$

Conditionals and Marginals of a Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

$$p(\mathbf{x}) = p\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}\right) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix} \,\middle|\, \begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right)$$

$$p(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_{11}) \qquad p(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_{22})$$

$$p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}\left(\mathbf{x}_1|\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\,,\; \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)$$

Probabilistic Discriminative Modeling

We observe some data $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\} = (\mathbf{X}, \mathbf{y})$, related by the unknown parameter $\boldsymbol{\theta}$ through some $p(y|\mathbf{x}, \boldsymbol{\theta})$

Joint distribution: $p(y, \mathbf{x}, \boldsymbol{\theta})$ or $p(\mathcal{D}, \boldsymbol{\theta}) = p(\mathbf{X}, \mathbf{y}, \boldsymbol{\theta})$

Parameter Prior: $p(\boldsymbol{\theta})$

Likelihood: $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$

Posterior: $p(\boldsymbol{\theta}|\mathcal{D}) = \dfrac{p(\mathcal{D}, \boldsymbol{\theta})}{p(\mathcal{D})} = \dfrac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} = \dfrac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{X})}$

Evidence: $p(\mathcal{D}) = \int p(\mathcal{D}, \theta)\, d\theta = \int p(\mathcal{D}|\theta)p(\theta)\, d\theta$

Posterior predictive distribution:

$$p(y^*|\mathbf{x}^*, \mathcal{D}) = \int p(y^*|\mathbf{x}^*, \theta)p(\theta|\mathcal{D})\, d\theta$$

Probabilistic Regression with Linear Gaussian Models

$$y = \mathbf{w}^T \mathbf{x} + z$$

$$p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma_z^2) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma_z^2 \mathbf{I}) \qquad p(\mathbf{w}|\mathbf{0}, \mathbf{V}_0) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{V}_0)$$

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \sigma_z^2, \mathbf{V}_0) \propto \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma_z^2 \mathbf{I}) \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{V}_0) = \mathcal{N}(\mathbf{w}|\mathbf{w}_N, \mathbf{V}_N)$$

$$\begin{aligned} \mathbf{w}_N &= \frac{1}{\sigma_z^2} \mathbf{V}_N \mathbf{X}^T \mathbf{y} \\ \mathbf{V}_N^{-1} &= \mathbf{V}_0^{-1} + \frac{1}{\sigma_z^2} \mathbf{X}^T \mathbf{X} \end{aligned}$$

Probabilistic Regression with Linear Gaussian Models

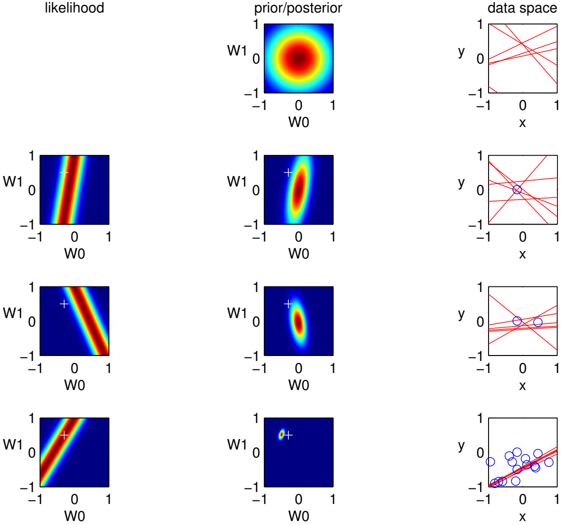


Figure: Source: Bishop's book

Predictive distribution

$$
\begin{aligned}
p(y^*|\mathbf{x}^*, \mathcal{D}, \sigma_z^2, \mathbf{V}_0) &= \int \mathcal{N}(y^*|\mathbf{w}^T\mathbf{x}^*, \sigma_z^2)\mathcal{N}(\mathbf{w}|\mathbf{w}_N, \mathbf{V}_N)\, d\mathbf{w} \\
&= \mathcal{N}(y^*|\mathbf{w}_N^T\mathbf{x}^*, \sigma_z^2 + \mathbf{x}^{*T}\mathbf{V}_N\mathbf{x}^*)
\end{aligned}
$$

- The predicted mean is the product between the posterior mean of the weights and the test input.
- The predictive uncertainty grow with the magnitude of the test input [1].

[1] Indeed, it grows with $\mathbf{x}^{*T}\mathbf{V}_N\mathbf{x}^*$, which is the Mahalanobis distance between $\mathbf{x}$ and $\mathbf{0}$ according to the eigenvectors of $\mathbf{V}_N$ (See Bishop Chapter 2.3)

Projections into Feature Space

Replace **x** by $\phi(\mathbf{x})$ and $\mathbf{X}$ by $\mathbf{\Phi}$ everywhere in the above equations.
If we design the right feature space ...

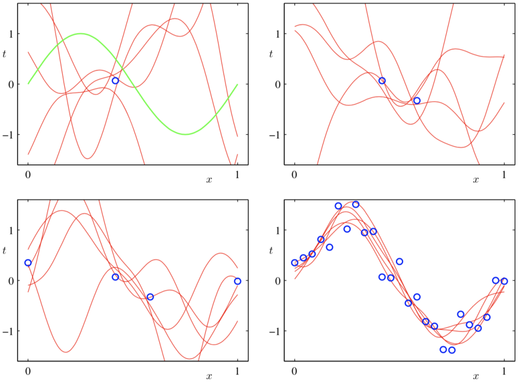


Figure: Source: Bishop's book

Marginal Likelihood (a.k.a. Model Evidence)

Useful for both Bayesian model comparison and hyperparmeter optimization

$$p(\mathbf{y}|\mathbf{\Phi}, \sigma_z^2, \mathbf{V}_0) = \int \mathcal{N}(\mathbf{y}|\mathbf{\Phi}\mathbf{w}, \sigma_z^2\mathbf{I})\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{V}_0)d\mathbf{w} = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_y)$$
$$= \frac{1}{(2\pi)^{N/2} \mid \mathbf{K}_y \mid^{1/2}} e^{-\frac{1}{2}\mathbf{y}^T\mathbf{K}_y^{-1}\mathbf{y}}$$

where[2] $\mathbf{K}_y = \frac{1}{\sigma_z}^2\mathbf{I} + \mathbf{\Phi}\mathbf{V}_0\mathbf{\Phi}^T$.

[2] Check out Page 93 of Bishop's book

Section 2

Gaussian Processes

Gaussian Processes

Instead of modeling $\mathbf{w}$ as Gaussian in a linear model $\mathbf{w}^T\mathbf{x} + z = f(\mathbf{x})$ we now model $f(\mathbf{x})$ as Gaussian

- $p(f|\mathbf{X}, \mathbf{y})$ is a distribution over functions
- A Gaussian Process (GP) assumes that $p(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N))$ is jointly Gaussian with mean $\boldsymbol{\mu}(\mathbf{X}) = [m(\mathbf{x}_1), \ldots m(\mathbf{x}_N)]^T$ and covariance $\boldsymbol{\Sigma}(\mathbf{X})$ given by $\Sigma_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$
- More formally

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

where

$$\begin{aligned} m(\mathbf{x}) &= E\{f(\mathbf{x})\} \\ k(\mathbf{x}, \mathbf{x}') &= E\{(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))^T\} \end{aligned}$$

Sampling from a Gaussian Process

$$f(\mathbf{x}) \sim GP(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'))$$

$$k(\mathbf{x}, \mathbf{x}') = \theta_0 e^{-\frac{\theta_1}{2}||\mathbf{x}-\mathbf{x}'||^2} + \theta_2 + \theta_3 \mathbf{x}^T \mathbf{x}'$$

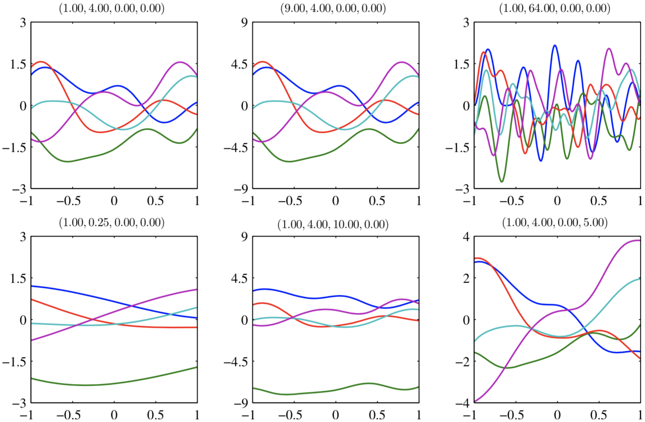
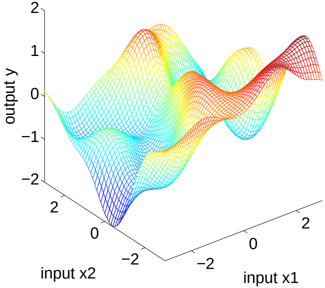
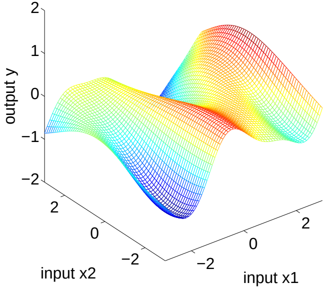


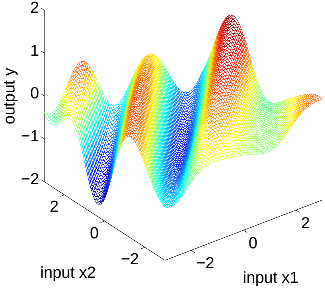
Figure: Source: Bishop's book

Sampling from a Gaussian Process

Source: Murphy's book







Predictions Using Noise-Free Observations

$$p\left(\begin{bmatrix}\mathbf{f}\\ \mathbf{f}^*\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{f}\\ \mathbf{f}^*\end{bmatrix} \;\Big|\; \begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}^*\end{bmatrix}, \begin{bmatrix}\mathbf{K} & \mathbf{K}^*\\ \mathbf{K}^{*T} & \mathbf{K}^{**}\end{bmatrix}\right)$$

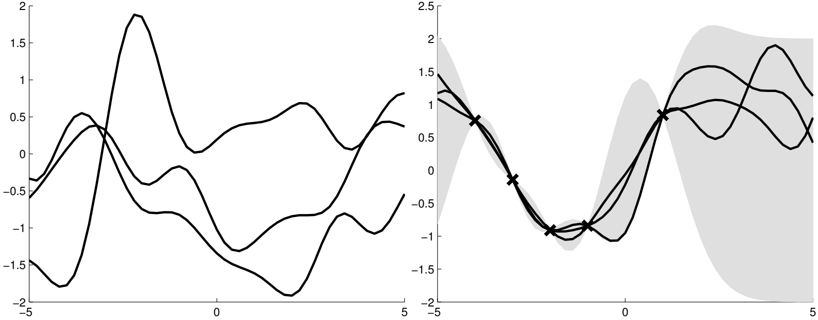
$$\begin{aligned} p(\mathbf{f}^*|\mathbf{X}^*, \mathbf{f}, \mathbf{X}) &= \mathcal{N}(\mathbf{f}^*|\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \\ \boldsymbol{\mu}_* &= \boldsymbol{\mu}(\mathbf{X}^*) + \mathbf{K}^{*T}\mathbf{K}^{-1}(\mathbf{f} - \boldsymbol{\mu}(\mathbf{X})) \\ \boldsymbol{\Sigma}_* &= \mathbf{K}^{**} - \mathbf{K}^{*T}\mathbf{K}^{-1}\mathbf{K}^* \end{aligned}$$

Zero mean prior

$$p(\mathbf{f}^*|\mathbf{X}^*, \mathbf{f}, \mathbf{X}) = \mathcal{N}(\mathbf{f}^*|\mathbf{K}^{*T}\mathbf{K}^{-1}\mathbf{f}, \mathbf{K}^{**} - \mathbf{K}^{*T}\mathbf{K}^{-1}\mathbf{K}^*)$$

Predictions Using Noise-Free Observations

Source: Murphy's book



$$k(\mathbf{x}, \mathbf{x}') = \theta_0 e^{-\frac{\theta_1}{2}||\mathbf{x}-\mathbf{x}'||^2}$$

Predictions Using Noisy Observations

$$y = f(\mathbf{x}) + z \qquad z \sim \mathcal{N}(z|0, \sigma_z^2)$$

$$\mathrm{Cov}(y_p, y_q) = k(\mathbf{x}_p, \mathbf{x}_q) + \sigma_z^2 \mathbb{I}(p = q) \quad \mathrm{Cov}(\mathbf{y}|\mathbf{X}) = \mathbf{K}_y = \mathbf{K} + \sigma_z^2 \mathbf{I}_N$$

$$p\left(\begin{bmatrix}\mathbf{y}\\ \mathbf{f}^*\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{y}\\ \mathbf{f}^*\end{bmatrix} \Big| \mathbf{0}, \begin{bmatrix}\mathbf{K}_y & \mathbf{K}^*\\ \mathbf{K}^{*T} & \mathbf{K}^{**}\end{bmatrix}\right)$$

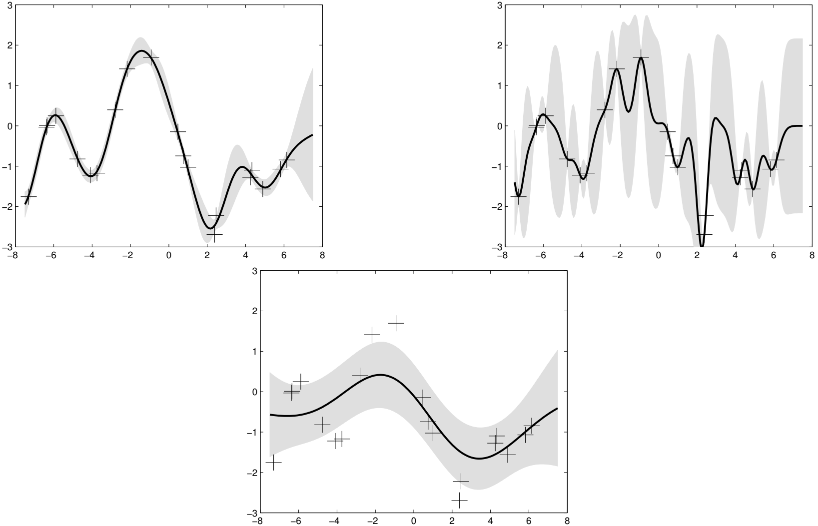
$$\begin{aligned} p(\mathbf{f}^*|\mathbf{X}^*, \mathbf{y}, \mathbf{X}) &= \mathcal{N}(\mathbf{f}^*|\mathbf{K}^{*T}\mathbf{K}_y^{-1}\mathbf{y}, \mathbf{K}^{**} - \mathbf{K}^{*T}\mathbf{K}_y^{-1}\mathbf{K}^*) \\ &= \mathcal{N}(\mathbf{f}^*|\mathbf{K}^{*T}(\mathbf{K} + \sigma_z^2\mathbf{I}_N)^{-1}\mathbf{y}, \mathbf{K}^{**} - \mathbf{K}^{*T}(\mathbf{K} + \sigma_z^2\mathbf{I}_N)^{-1}\mathbf{K}^*) \end{aligned}$$

Single value prediction:

$$p(f^*|\mathbf{x}^*, \mathbf{y}, \mathbf{X}) \;=\; \mathcal{N}(f^*|\mathbf{k}^{*T}\mathbf{K}_y^{-1}\mathbf{y}, \mathbf{k}^{**} - \mathbf{k}^{*T}\mathbf{K}_y^{-1}\mathbf{k}^*)$$

Predictions Using Noisy Observations

$$k(\mathbf{x}, \mathbf{x}') = \theta_0 e^{-\frac{\theta_1}{2}||\mathbf{x}-\mathbf{x}'||^2}$$



Source: Murphy's book

Estimating the Kernel Parameters

Instead of exhaustive search by CV, we can optimize again the marginal likelihood or evidence

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{X}) &= \int p(\mathbf{y}|\mathbf{f},\mathbf{X})p(\mathbf{f}|\mathbf{X})d\mathbf{f} = \mathcal{N}(\mathbf{0},\mathbf{K}_y) \\
&= \frac{1}{(2\pi)^{N/2} \mid \mathbf{K}_y \mid^{1/2}} e^{-\frac{1}{2}\mathbf{y}^T\mathbf{K}_y^{-1}\mathbf{y}}
\end{aligned}
$$

For a kernel parameter θ

$$
\begin{aligned}
\frac{\partial}{\partial\theta}\log p(\mathbf{y}|\mathbf{X}) &= \frac{1}{2}\mathbf{y}^T\mathbf{K}_y^{-1}\frac{\partial\mathbf{K}_y}{\partial\theta}\mathbf{K}_y^{-1}\mathbf{y} - \frac{1}{2}\operatorname{tr}\left(\mathbf{K}_y^{-1}\frac{\partial\mathbf{K}_y}{\partial\theta}\right) \\
&= \frac{1}{2}\operatorname{tr}\left((\boldsymbol{\alpha}\boldsymbol{\alpha}^T - \mathbf{K}_y^{-1})\frac{\partial\mathbf{K}_y}{\partial\theta}\right)
\end{aligned}
$$

where $\boldsymbol{\alpha} = \mathbf{K}_y^{-1}\mathbf{y}$. It takes $\mathcal{O}(N^3)$ time to compute $\mathbf{K}_y^{-1}$ and $\mathcal{O}(N^2)$ time per hyperparameter to compute the gradient.

Section 3

GPs for Classification

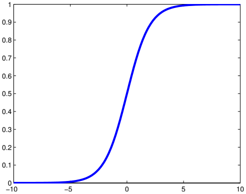
Binary GP Classifier

The simplest way to obtain a binary classifier is to employ the same “trick” of logistic or probit regression:

$$p(y = +1|\mathbf{x}) = \sigma(f(\mathbf{x}))$$

where $f(\mathbf{x})$ is a GP and $\sigma(\cdot)$ is the logistic function

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^{x}}{e^{x} + 1}$$



Binary GP Classifier

$$p(y = +1|\mathbf{x}) = \sigma(f(\mathbf{x}))$$

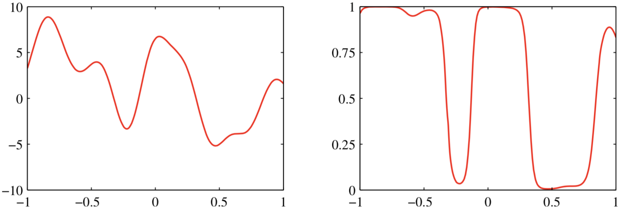


Figure: Source: Bishop's book

$$p(y = +1|\mathbf{x}) = \sigma(f(\mathbf{x}))$$

- $f(\mathbf{x})$ is a GP.
- Our goal is to determine the predictive distribution $p(y^* = 1|\mathbf{y}, \mathbf{X})$:

$$p(y^* = 1|\mathbf{f}) = \int p(y^* = 1|f(\mathbf{x}^*))p(f(\mathbf{x}^*)|\mathbf{y})df(\mathbf{x}^*)$$

- This integral is analytically intractable.
- Find a Gaussian approximation to $p(f(\mathbf{x}^*)|\mathbf{y})$ (Laplace's method or Expectation Propagation).