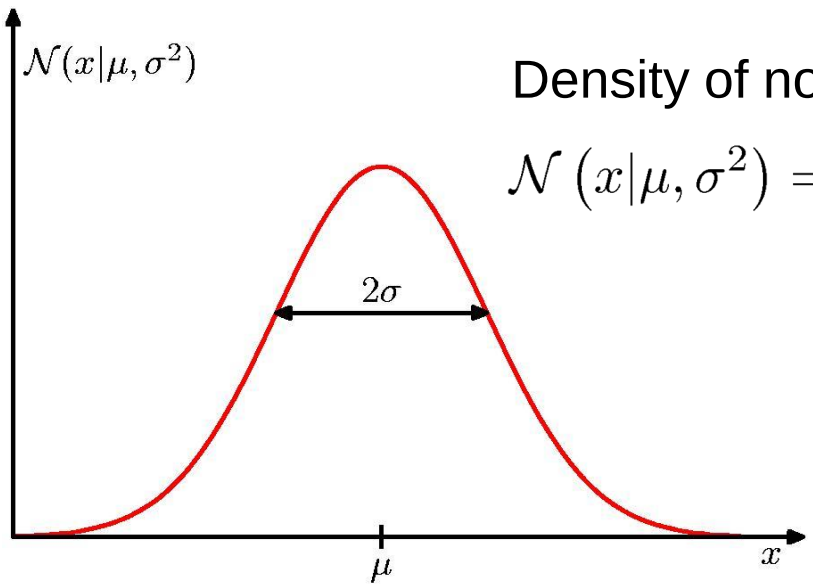
Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Bayesian Learning

Tobias Scheffer, Niels Landwehr

Remember: Normal Distribution

- Distribution over $x \in \mathbb{R}$.
- Density function with parameters $\mu \in \mathbb{R}$ (mean) and $\sigma^2 \in \mathbb{R}$ (variance).



Density of normal distribution

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

Remember: Multivariate Normal Distribution

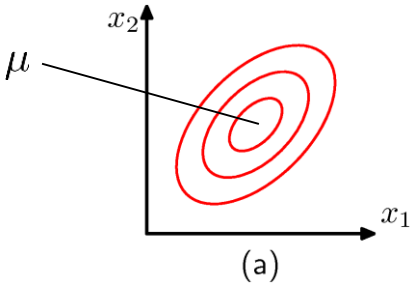
- Distribution over vectors $\mathbf{x} \in \mathbb{R}^D$.
- Density function with parameters $\boldsymbol{\mu} \in \mathbb{R}^D, \boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$.

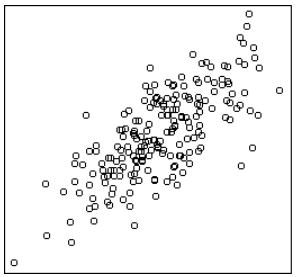
mean vector

covariance matrix

$$\mathrm{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi^{D/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

- Example $D=2$: density, sample from distribution





Overview

- Basic concepts of Bayesian learning
- Linear regression:
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Posterior distribution over models,
  - Bayesian prediction, predictive distribution,
- Linear classification: logistic regression.
  - Predictive distribution,
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Bayesian Prediction.
- Naive Bayesian classifier.
- Nonlinear models: Gaussian processes.

Statistics & Machine Learning

- Machine learning: tightly related to inductive statistics.
- Two areas in Statistics:
  - *Descriptive Statistics*: description and examination of the properties of data.

Mean values

Variances

Difference between Populations

  - *Inductive Statistics*: What conclusions can be drawn from data about the underlying reality?

Model building

Explanations for observations

Relationships and patterns in the data

Frequentist vs. Bayesian Probabilities

- Frequentist probabilities
  - Describe the possibility of an occurrence of an intrinsically stochastic event (e.g., a coin toss).
  - Defined as limits of *relative frequencies* of possible outcomes in a *repeatable experiment*

"If one throws a fair coin 1000 times, it will land on heads about 500 times"

"In 1 gram of Potassium-40, around 260,000 nuclei decay per second"

Frequentist vs. Bayesian Probabilities

- Bayesian “subjective“ probabilities
  - Here, the reason for uncertainty is attributed to a lack of information.
  - How likely is it that suspect X killed the victim?
  - New Information (e.g., finger prints) can change these subjective probabilities.

- Bayesian view is more important in machine learning
- Frequentist and Bayesian perspectives are mathematically equivalent; in Bayesian statistics, probabilities are just used to model different things (lack of information).

Bayesian Statistics

- 1702-1761
- “An essay towards solving a problem in the doctrine of chances“, published in 1764.

- The work of Bayes laid the foundations for inductive Statistics.

- “Bayesian Probabilities“ offer an important perspective on uncertainty and probability.

Bayesian Probability in Machine Learning

- Model building: find an explanation for observations.
- What is the “most likely“ model? Trade-off between
  - Prior knowledge (a priori distribution over models),
  - Evidence (data, observations).
- Bayesian Perspective:
  - Evidence (data) changes the “subjective“ probability for models (explanation),
  - A posteriori model probability, MAP hypothesis.

Bayesian Model of Learning

System parameter θ^* ← Nature randomly determines θ^* according to $P(\theta)$

↓

Observations $\mathbf{y}$ for input $\mathbf{X}$

- "Nature" conducts random experiment, determines θ^*.
- System with parameter θ^* generates observations $\mathbf{y} = f_{\theta^*}(\mathbf{X})$.
- Bayesian inference inverts this process:
  - Given these assumptions about how $\mathbf{y}$ is generated,
  - Given the observed values of $\mathbf{y}$ for input matrix $\mathbf{X}$,
  - What is the most likely true value of θ?
  - What is the most likely value y^* for a new input $\mathbf{x}^*$?

Bayesian Model of Learning

System parameter θ^* ← Nature randomly determines θ^* according to $P(\theta)$

↓

Observations $\mathbf{y}$ for input $\mathbf{X}$

Likelihood of observing $\mathbf{y}|\mathbf{X}$ when model parameter is θ.

A priori ("prior") probability of nature choosing θ

- Bayes' equation:

$$P(\theta|\mathbf{X},\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{X},\theta)P(\theta)}{P(\mathbf{y}|\mathbf{X})}$$

Probability of observing $\mathbf{y}|\mathbf{X}$; independent of θ.

A posteriori ("posterior") probability that θ is the correct parameter given observations $\mathbf{y}|\mathbf{X}$.

Bayesian Model of Learning

- Maximum-likelihood (ML) model:
  - $\theta_{ML} = \arg\max_{\theta} P(\mathbf{y}|\mathbf{X}, \theta)$.

Likelihood

- Maximum-a-positeriori (MAP) model:
  - $\theta_{MAP} = \arg\max_{\theta} P(\theta|\mathbf{y}, \mathbf{X})$

A posteriori ("posterior") distribution

Bayesian Model of Learning

- Maximum-likelihood (ML) model:
  - $\theta_{ML} = \arg\max_{\theta} P(\mathbf{y}|\mathbf{X}, \theta)$.
- Maximum-a-positeriori (MAP) model:
  - $\theta_{MAP} = \arg\max_{\theta} P(\theta|\mathbf{y}, \mathbf{X}) = \arg\max_{\theta} \frac{P(\mathbf{y}|\mathbf{X}, \theta)P(\theta)}{P(\mathbf{y}|\mathbf{X})}$
    $= \arg\max_{\theta} P(\mathbf{y}|\mathbf{X}, \theta)P(\theta)$

Posterior ∝ likelihood x prior

Bayesian Model of Learning

- Maximum-likelihood (ML) model:
  - $\theta_{ML} = \arg\max_{\theta} P(\mathbf{y}|\mathbf{X}, \theta)$.
- Maximum-a-positeriori (MAP) model:
  - $\theta_{MAP} = \arg\max_{\theta} P(\theta|\mathbf{y}, \mathbf{X}) = \arg\max_{\theta} \frac{P(\mathbf{y}|\mathbf{X}, \theta)P(\theta)}{P(\mathbf{y}|\mathbf{X})}$
    $= \arg\max_{\theta} P(\mathbf{y}|\mathbf{X}, \theta)P(\theta)$
- Most likely value $\mathbf{y}^*$ for new input $\mathbf{x}^*$ (Bayes-optimal decision):
  - $\mathbf{y}^* = \arg\max_{y} P(y|\mathbf{x}^*, \mathbf{y}, \mathbf{X})$

Bayesian Model of Learning

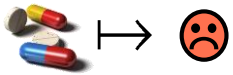
- Maximum-likelihood (ML) model:
  - $\theta_{ML} = \arg\max_{\theta} P(\mathbf{y}|\mathbf{X}, \theta)$.
- Maximum-a-positeriori (MAP) model:
  - $\theta_{MAP} = \arg\max_{\theta} P(\theta|\mathbf{y}, \mathbf{X}) = \arg\max_{\theta} \frac{P(\mathbf{y}|\mathbf{X}, \theta)P(\theta)}{P(\mathbf{y}|\mathbf{X})}$
    $= \arg\max_{\theta} P(\mathbf{y}|\mathbf{X}, \theta)P(\theta)$
- Most likely value $\mathbf{y}^*$ for new input $\mathbf{x}^*$ (Bayes-optimal decision):
  - $\mathbf{y}^* = \arg\max_{y} P(y|\mathbf{x}^*, \mathbf{y}, \mathbf{X})$
  - $P(y^*|\mathbf{x}^*, \mathbf{y}, \mathbf{X}) = \int P(y^*, \theta|\mathbf{x}^*, \mathbf{y}, \mathbf{X})d\theta$
    $= \int P(y^*|\mathbf{x}^*, \theta)P(\theta|\mathbf{y}, \mathbf{X})d\theta$

Predictive distribution

"Bayesian model averaging". Often computationally infeasible, but has a closed-form solution in some cases.

Linear Regression Models

- Training data:
  - $\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$ 
  - $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ 
- Model
  - $f_{\boldsymbol{\theta}} : X \to Y$

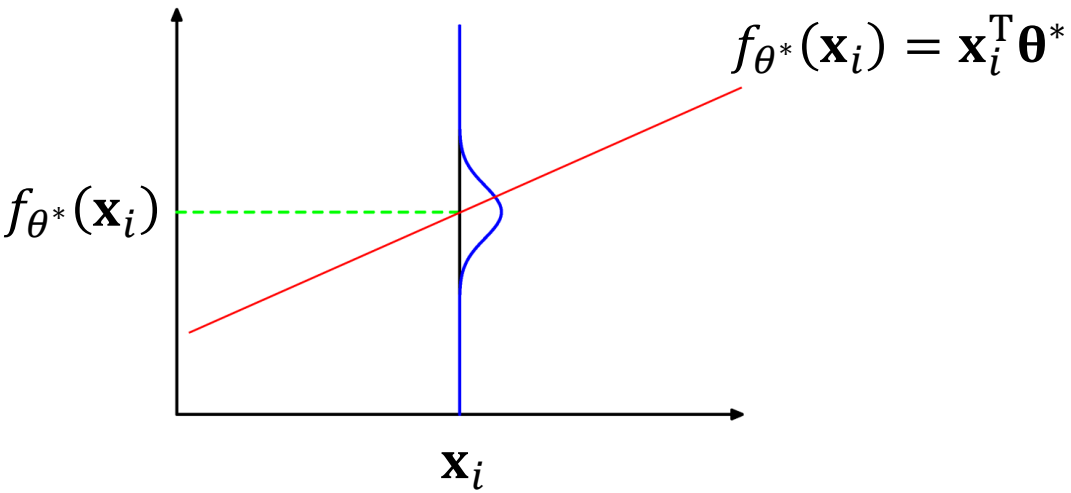
    
  - $f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}$

Overview

- Basic concepts of Bayesian learning
- Linear regression:
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Posterior distribution over models,
  - Bayesian prediction, predictive distribution,
- Linear classification: logistic regression.
  - Predictive distribution,
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Bayesian Prediction.
- Naive Bayesian classifier.
- Nonlinear models: Gaussian processes.

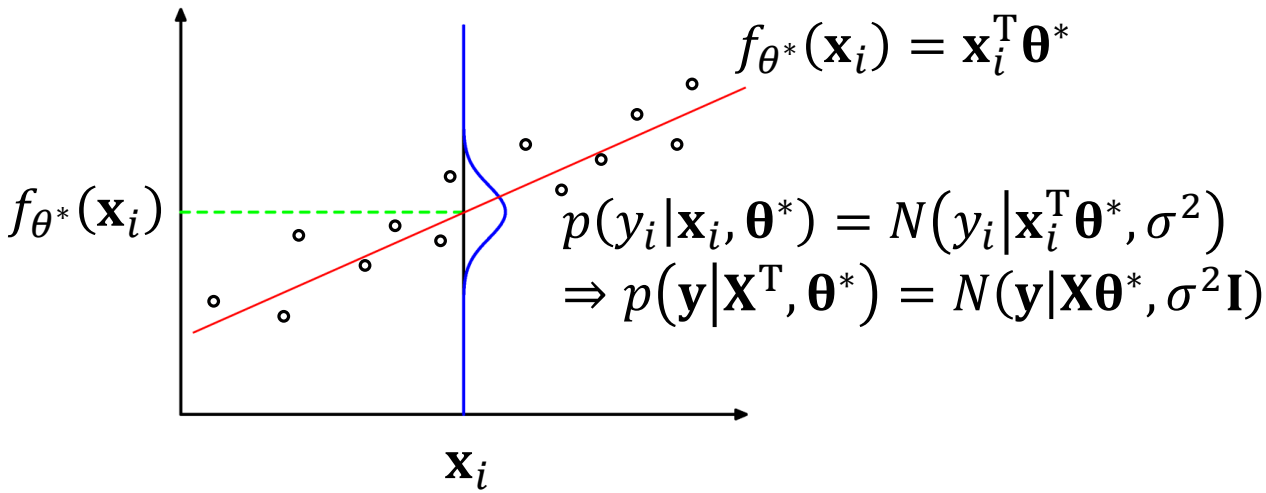
Probabilistic Linear Regression

- Assumption 1: Nature generates parameter $\boldsymbol{\theta}^*$ of a linear function $f_{\boldsymbol{\theta}^*}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}^*$ according to $p(\boldsymbol{\theta})$.



Probabilistic Linear Regression

- Assumption 1: Nature generates parameter $\boldsymbol{\theta}^*$ of a linear function $f_{\boldsymbol{\theta}^*}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}^*$ according to $p(\boldsymbol{\theta})$.
- Assumption 2: Given inputs $\mathbf{X}$, nature generates outputs $\mathbf{y}$:
  - $y_i = f_{\boldsymbol{\theta}^*}(\mathbf{x}_i) + \epsilon_i$ with $\epsilon_i \sim N(\epsilon|0, \sigma^2)$.
  - $p(y_i|\mathbf{x}_i, \boldsymbol{\theta}^*) = N(y_i|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}^*, \sigma^2)$



- In reality, we have $\mathbf{y}, \mathbf{X}$ and want to make inferences about $\boldsymbol{\theta}$.

Maximum Likelihood for Linear Regression

- Maximum-likelihood (ML) model:

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} N(y_i|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}, \sigma^2)$$

Training instances are independent

Maximum Likelihood for Linear Regression

- Maximum-likelihood (ML) model:

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} N(y_i|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}, \sigma^2)$$

$$= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}\right)^2\right\}$$

Maximum Likelihood for Linear Regression

- Maximum-likelihood (ML) model:

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} N(y_i|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}, \sigma^2)$$

$$= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}\right)^2\right\}$$

- Log is a mononomic transformation:
  - $\arg\max_{\theta} P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \arg\max_{\theta} \log P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$
- Also constant terms (constant in $\boldsymbol{\theta}$) can be dropped.

Maximum Likelihood for Linear Regression

- Maximum-likelihood (ML) model:

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} P(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} N(y_i|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta},\sigma^2)$$

$$= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}\right)^2\right\}$$

$$\boxed{= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}\right)^2}$$

Unregularized linear regression with squared loss

$\log \prod e^x = \sum x$

- Log is a monotonic transformation:
  - $\arg\max_{\theta} P(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = \arg\max_{\theta} \log P(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})$
- Also constant terms (constant in $\boldsymbol{\theta}$) can be dropped

Maximum Likelihood for Linear Regression

- Maximum-likelihood (ML) model:

$$\boldsymbol{\theta}_{ML} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}\right)^2$$

  - Known as least-squares method in statistics.

- Setting the derivative to zero gives closed-form solution:

$$\boldsymbol{\theta}_{ML} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- Inversion of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is numerically unstable.

Maximum Likelihood for Linear Regression

- The maximum-likelihood model is only based on the data, it is independent of any prior knowledge or domain assumptions.
- Calculating the maximum-likelihood model is numerically unstable.
- Regularized least squares works much better in practice.

Overview

- Basic concepts of Bayesian learning
- Linear regression:
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Posterior distribution over models,
  - Bayesian prediction, predictive distribution,
- Linear classification: logistic regression.
  - Predictive distribution,
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Bayesian Prediction.
- Naive Bayesian classifier.
- Nonlinear models: Gaussian processes.

Overview

- Basic concepts of Bayesian learning
- Linear regression:
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Posterior distribution over models,
  - Bayesian prediction, predictive distribution,
- Linear classification (logistic regression):
  - Predictive distribution,
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Bayesian Prediction.
- Nonlinear models: Gaussian processes.

MAP for Linear Regression

- Maximum-a-positeriori (MAP) model:
  - $\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \arg\max_{\boldsymbol{\theta}} \frac{P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{y}|\mathbf{X})}$

MAP for Linear Regression

- Maximum-a-positeriori (MAP) model:
  - $$\begin{aligned}\boldsymbol{\theta}_{MAP} &= \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = \arg\max_{\boldsymbol{\theta}} \frac{P(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{y}|\mathbf{X})} \\ &= \arg\max_{\boldsymbol{\theta}} P(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})P(\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \log\big(P(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})P(\boldsymbol{\theta})\big) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i,\boldsymbol{\theta}) + \log P(\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log N\big(y_i|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta},\sigma^2\big) + \log P(\boldsymbol{\theta}) \\ &= \cdots\end{aligned}$$

Training instances are independent

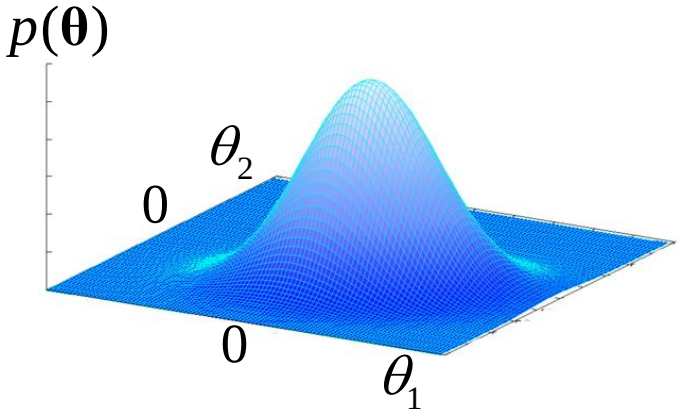
MAP for Linear Regression: Prior

- Nature generates parameter $\boldsymbol{\theta}^*$ of linear model $f_{\boldsymbol{\theta}^*}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}^*$ according to $p(\boldsymbol{\theta})$.
- For convenience, assume $p(\boldsymbol{\theta}) = \mathrm{N}(\boldsymbol{\theta}|\mathbf{0}, \sigma_p^2\mathbf{I})$.

$$p(\boldsymbol{\theta}) = \mathrm{N}\,(\boldsymbol{\theta}\,|\,\mathbf{0}, \sigma_p^2\mathbf{I})$$

$$= \frac{1}{2\pi^{m/2}\sigma_p}\exp\left(-\frac{1}{2\sigma_p^2}|\boldsymbol{\theta}|^2\right)$$

$\sigma_p^2 \in \mathbb{R}$ controls strength of prior



MAP for Linear Regression

- Maximum-a-positeriori (MAP) model:
  - $\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log N\left(y_i \middle| \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}, \sigma^2\right) + \log P(\boldsymbol{\theta})$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}\right)^2 + \log \frac{1}{2\pi^{\frac{m}{2}}\sigma_p} - \frac{1}{2\sigma_p^2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}$$

All terms that are constant in θ can be dropped.

MAP for Linear Regression

- Maximum-a-positeriori (MAP) model:
  - $\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log N\left(y_i \middle| \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}, \sigma^2\right) + \log P(\boldsymbol{\theta})$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}\right)^2 + \log \frac{1}{2\pi^{\frac{m}{2}}\sigma_p} - \frac{1}{2\sigma_p^2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}$$

$$= \boxed{\operatorname{argmin}_{\theta} \sum_{i=1}^{n} \frac{1}{2\sigma^2}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}\right)^2 + \frac{1}{2\sigma_p^2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}}$$

ℓ_2-regularized linear regression with squared loss (ridge regression).

MAP for Linear Regression

- Maximum-a-positeriori (MAP) model:
  - $\boldsymbol{\theta}_{MAP} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}\right)^2 - \underbrace{\frac{\sigma^2}{\sigma_p^2}}_{\lambda} \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}$

- Same optimization criterion as ridge regression.
- Analytic solution (see lecture on ridge regression):
  - $\boldsymbol{\theta}_{MAP} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y}$

Overview

- Basic concepts of Bayesian learning
- Linear regression:
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Posterior distribution over models,
  - Bayesian prediction, predictive distribution,
- Linear classification: logistic regression).
  - Predictive distribution,
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Bayesian Prediction.
- Naive Bayesian classifier.
- Nonlinear models: Gaussian processes.

Posterior for Linear Regression

- Posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{y}, \mathbf{X}$:
  - $P(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \frac{P(\mathbf{y}|\mathbf{X}, \theta)P(\theta)}{P(\mathbf{y}|\mathbf{X})} = \frac{1}{Z} P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})P(\boldsymbol{\theta})$

Posterior for Linear Regression

- Posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{y}, \mathbf{X}$:
  - $P(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = \frac{P(\mathbf{y}|\mathbf{X},\theta)P(\theta)}{P(\mathbf{y}|\mathbf{X})} = \frac{1}{Z}P(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})P(\boldsymbol{\theta})$

$$= \frac{1}{Z}N\left(\mathbf{y}\middle|\mathbf{X}^{\mathrm{T}}\boldsymbol{\theta},\sigma^2\mathbf{I}\right)N\left(\boldsymbol{\theta}\middle|\mathbf{0},\sigma_p^2\,\mathbf{I}\right)$$

$$= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta})^2\right\}$$

The normal distribution is the conjugate of itself. Therefore $N(\cdot \mid \cdot,\cdot)N(\cdot \mid \cdot,\cdot) = N(\cdot \mid \cdot,\cdot)$

Posterior for Linear Regression

- Posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{y}, \mathbf{X}$:
  - $P(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = \frac{P(\mathbf{y}|\mathbf{X},\theta)P(\theta)}{P(\mathbf{y}|\mathbf{X})} = \frac{1}{Z}P(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})P(\boldsymbol{\theta})$
    $= \frac{1}{Z}N\left(\mathbf{y}\middle|\mathbf{X}^{\mathrm{T}}\boldsymbol{\theta},\sigma^2\mathbf{I}\right)N\left(\boldsymbol{\theta}\middle|\mathbf{0},\sigma_p^2\,\mathbf{I}\right)$
    $= N(\boldsymbol{\theta}|\overline{\boldsymbol{\theta}},\mathbf{A}^{-1})$
  - With $\overline{\boldsymbol{\theta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$
  - And $\mathbf{A} = \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \sigma_p^{-2}\mathbf{I}$.

Mean value of posterior: $\boldsymbol{\theta}_{MAP}$

Example MAP solution regression

- Training data:

$$\mathbf{x}_1 = \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix}, \qquad \mathbf{x}_2 = \begin{pmatrix} 4 \\ 3 \\ 2 \end{pmatrix}, \qquad \mathbf{x}_3 = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix},$$

$$y_1 = 2 \qquad y_2 = 3 \qquad y_3 = 4$$

- Matrix notation (adding constant attribute):

$$\mathbf{X} = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 1 & 4 & 3 & 2 \\ 1 & 0 & 1 & 2 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$$

Example MAP solution regression

- Choose
  - Variance of prior: $\sigma_p = 1$
  - Noise parameter: $\sigma = 0.5$

- Compute: $\bar{\boldsymbol{\theta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

$$\bar{\boldsymbol{\theta}} = \left(\begin{pmatrix} 1 & 2 & 3 & 0 \\ 1 & 4 & 3 & 2 \\ 1 & 0 & 1 & 2 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} 1 & 2 & 3 & 0 \\ 1 & 4 & 3 & 2 \\ 1 & 0 & 1 & 2 \end{pmatrix} + 0.25 \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 2 & 3 & 0 \\ 1 & 4 & 3 & 2 \\ 1 & 0 & 1 & 2 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$$

$$\approx \begin{pmatrix} 0.7975 \\ -0.5598 \\ 0.7543 \\ 1.1217 \end{pmatrix}$$

Example MAP solution regression

- Predictions of model $\bar{\boldsymbol{\theta}}$ on the training data:

$$\hat{\mathbf{y}} = \mathbf{X}\bar{\boldsymbol{\theta}} = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 1 & 4 & 3 & 2 \\ 1 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 0.7975 \\ -0.5598 \\ 0.7543 \\ 1.1217 \end{pmatrix} = \begin{pmatrix} 1.9408 \\ 3.0646 \\ 3.7952 \end{pmatrix}$$

Posterior and Regularized Loss Function

- MAP model:
  - $$\begin{aligned}\boldsymbol{\theta}_{MAP} &= \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) = \arg\max_{\boldsymbol{\theta}} \frac{P(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{y}|\mathbf{X})} \\ &= \arg\max_{\boldsymbol{\theta}} P(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})P(\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \log P(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) + \log P(\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i,\boldsymbol{\theta}) + \log P(\boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}\right)^2 + \frac{\sigma^2}{\sigma_p^2}\boldsymbol{\theta}^{\mathbf{T}}\boldsymbol{\theta}\end{aligned}$$

Likelihood ≈ loss function

Prior ≈ regularizer

Sequential Learning

- Training examples arrive sequentially.
- Each training example $(\mathbf{x}_i, y_i)$ changes prior $p_{i-1}(\boldsymbol{\theta})$ into posterior $p_{i-1}(\boldsymbol{\theta}|y_i, \mathbf{x}_i)$ which becomes the new prior $p_i(\boldsymbol{\theta})$
  - $P(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \frac{1}{Z} P_0(\boldsymbol{\theta}) P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$

$$= \frac{1}{Z} P_0(\boldsymbol{\theta}) \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \boldsymbol{\theta})$$

$$= \frac{1}{Z} \underbrace{\underbrace{\underbrace{P_0(\boldsymbol{\theta}) P(y_1|\mathbf{x}_1, \boldsymbol{\theta})}_{P_1(\boldsymbol{\theta})} P(y_2|\mathbf{x}_2, \boldsymbol{\theta})}_{P_2(\boldsymbol{\theta})} P(y_3|\mathbf{x}_3, \boldsymbol{\theta})}_{P_3(\boldsymbol{\theta})} \ldots P(y_n|\mathbf{x}_n, \boldsymbol{\theta})$$

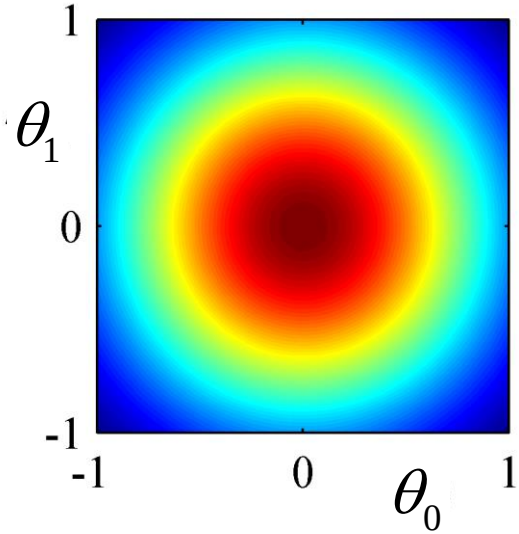
Example: Sequential Learning

[from Chris Bishop, Pattern Recognition and Machine Learning]

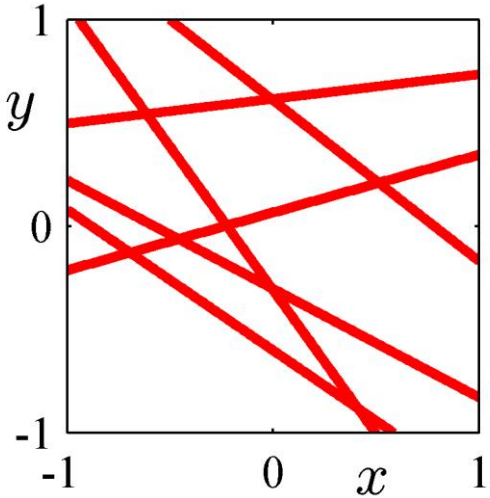
$f_\theta(x) = \theta_0 + \theta_1 x$ (one-dimensional regression)

Sequential update: $P_0(\boldsymbol{\theta})$

$P_0(\boldsymbol{\theta})$



Sample from $P_0(\boldsymbol{\theta})$

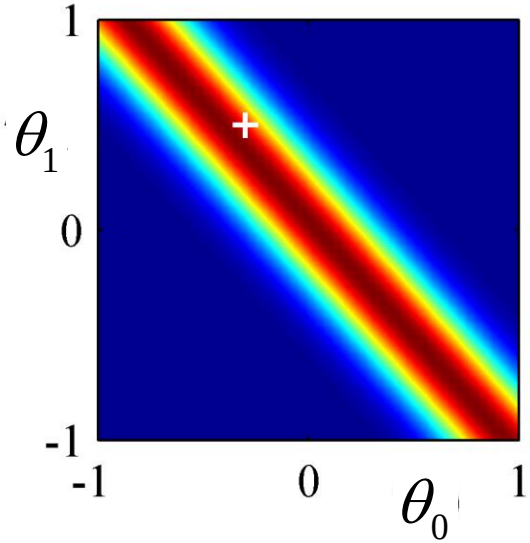


Example: Sequential Learning

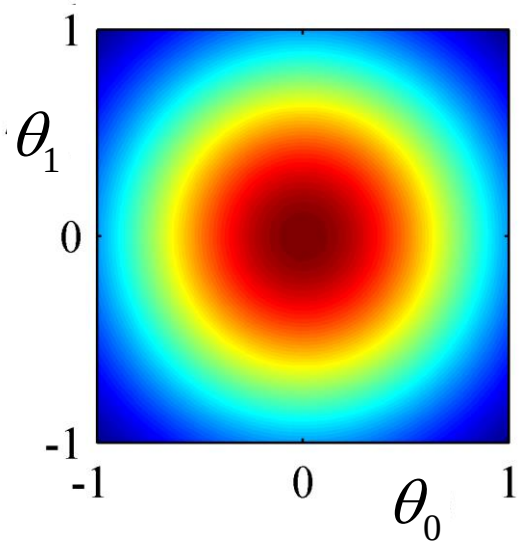
$f_\theta(x) = \theta_0 + \theta_1 x$ (one-dimensional regression)

Sequential update: $P_0(\boldsymbol{\theta})$

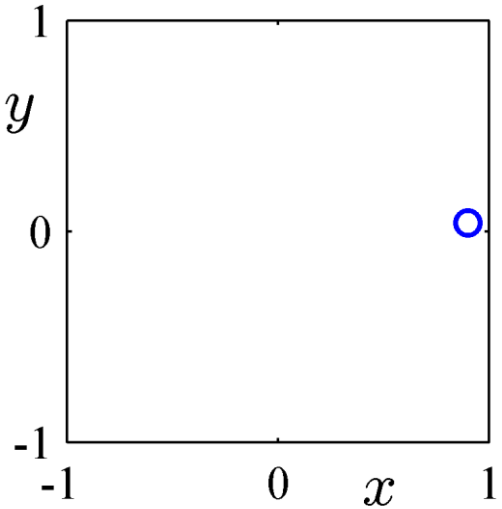
Likelihood $P(y_1|x_1, \boldsymbol{\theta})$



$P_0(\boldsymbol{\theta})$



Instance x_1, y_1

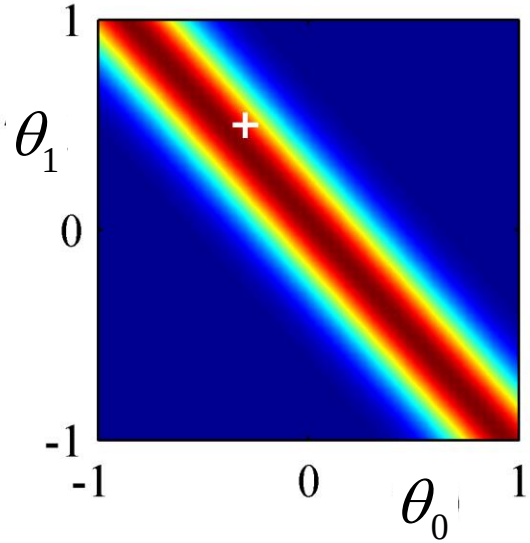


Example: Sequential Learning

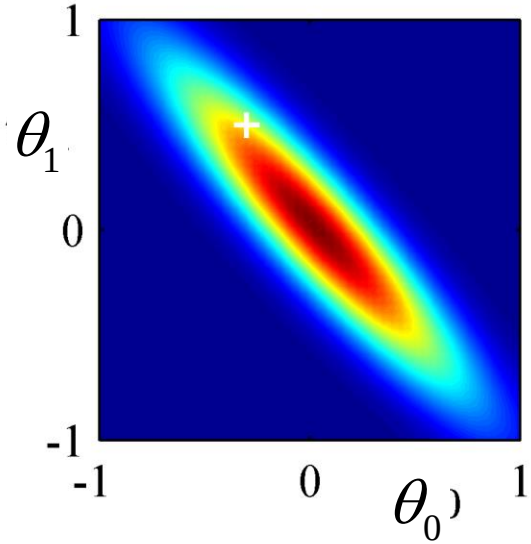
$f_\theta(x) = \theta_0 + \theta_1 x$ (one-dimensional regression)

Sequential update: $P_1(\boldsymbol{\theta}) \propto P_0(\boldsymbol{\theta}) P(y_1 | x_1, \boldsymbol{\theta})$

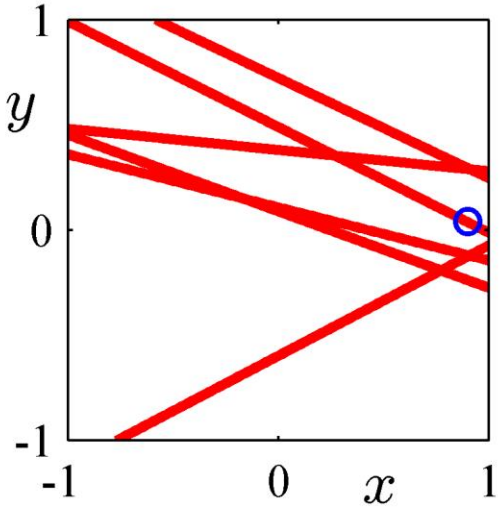
Likelihood $P(y_1 | x_1, \boldsymbol{\theta})$



$P_1(\boldsymbol{\theta})$



Sample from $P_1(\boldsymbol{\theta})$

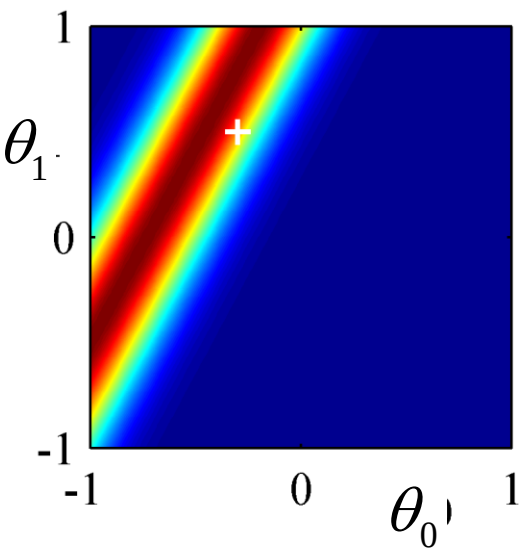


Example: Sequential Learning

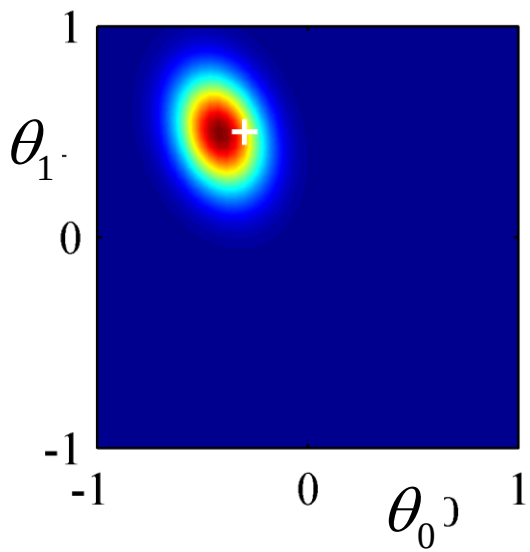
$f_\theta(x) = \theta_0 + \theta_1 x$ (one-dimensional regression)

Sequential update: $P_2(\boldsymbol{\theta}) \propto P_1(\boldsymbol{\theta}) P(y_2 | x_2, \boldsymbol{\theta})$

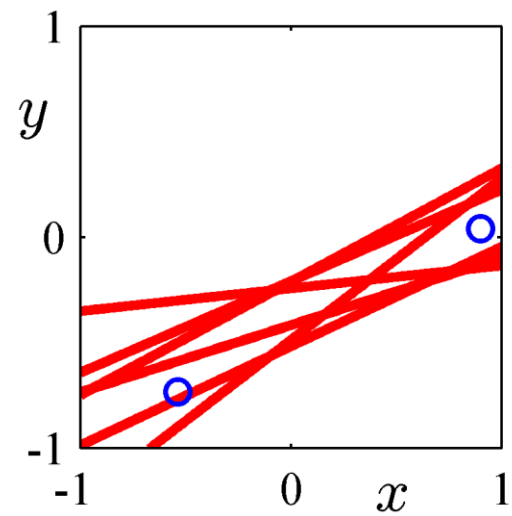
Likelihood $P(y_2 | x_2, \boldsymbol{\theta})$



$P_2(\boldsymbol{\theta})$



Sample from $P_2(\boldsymbol{\theta})$

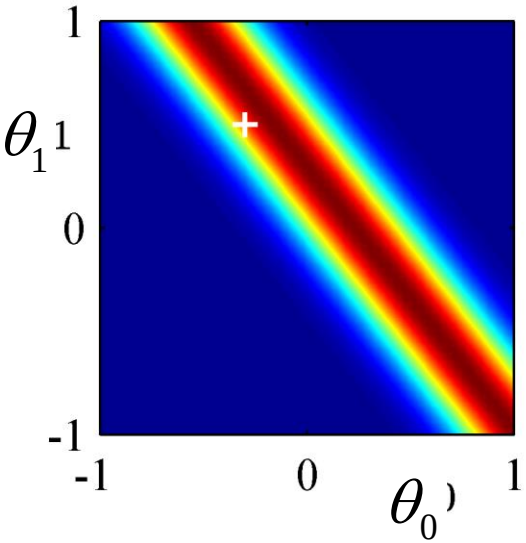


Example: Sequential Learning

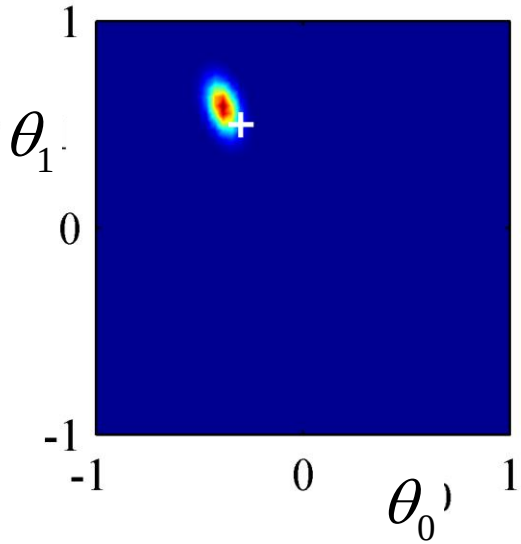
$f_\theta(x) = \theta_0 + \theta_1 x$ (one-dimensional regression)

Sequential update: $P_n(\boldsymbol{\theta}) \propto P_{n-1}(\boldsymbol{\theta}) P(y_n | x_n, \boldsymbol{\theta})$

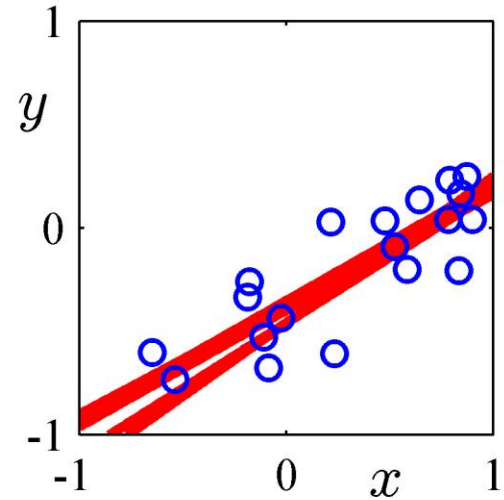
Likelihood $P(y_n | x_n, \boldsymbol{\theta})$



$P_n(\boldsymbol{\theta})$



Sample from $P_n(\boldsymbol{\theta})$



Learning and Prediction

- So far, we have always separated *learning* from *prediction*.
- Learning:
  - $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \hat{R}(\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}) + \Omega(\boldsymbol{\theta})$
- Prediction:
  - $y^* = f_{\boldsymbol{\theta}^*}(\mathbf{x}^*)$

- For instance, in MAP linear regression, *learning* is
  - $\boldsymbol{\theta}_{MAP} = \arg\max_{\theta} P(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$.
- And prediction is
  - $y^* = \boldsymbol{\theta}_{MAP}^T \mathbf{x}^*$.

Learning and Prediction

- So far, we have always separated *learning* from *prediction*.
- And there are good reasons to do this:
  - Learning can require processing massive amounts of training data which can take a long time.
  - Predictions may have to be made in real time.

- However, sometimes, when relatively few data are available and an accurate prediction is worth waiting for, one can directly search for the best possible prediction:
  - $\mathbf{y}^* = \arg\max_y P(y|\mathbf{x}^*, \mathbf{y}, \mathbf{X})$
  - Most likely $\mathbf{y}^*$ for new input $\mathbf{x}^*$ given training data $\mathbf{y}, \mathbf{X}$.

Overview

- Basic concepts of Bayesian learning
- Linear regression:
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Posterior distribution over models,
  - Bayesian prediction, predictive distribution,
- Linear classification: logistic regression).
  - Predictive distribution,
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Bayesian Prediction.
- Naive Bayesian classifier.
- Nonlinear models: Gaussian processes.

Bayes-Optimal Prediction

- Bayes-optimal decision: most likely value $\mathbf{y}^*$ for new input $\mathbf{x}^*$
  - $y^* = \arg\max_y P(y|\mathbf{x}^*, \mathbf{y}, \mathbf{X})$
- Predictive distribution for new input $\mathbf{x}^*$ given training data:
  - $$\begin{aligned} P(y|\mathbf{x}^*, \mathbf{y}, \mathbf{X}) &= \int P(y, \boldsymbol{\theta}|\mathbf{x}^*, \mathbf{y}, \mathbf{X}) d\boldsymbol{\theta} \\ &= \int P(y|\boldsymbol{\theta}, \mathbf{x}^*, \mathbf{y}, \mathbf{X}) P(\boldsymbol{\theta}|\mathbf{x}^*, \mathbf{y}, \mathbf{X}) \mathrm{d}\boldsymbol{\theta} \\ &= \int P(y|\boldsymbol{\theta}, \mathbf{x}^*) P(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) \mathrm{d}\boldsymbol{\theta} \end{aligned}$$

Sum rule

Product rule

Bayesian model averaging

Independence assumptions from model data generation process

- Bayes-optimal decision is made by a weighted sum over all values of the model parameters.
- In general, there is no single model $\boldsymbol{\theta}^*$ in the model space that always makes the Bayes-optimal decision.

Bayes-Optimal Prediction

- Bayes-optimal decision is made by a weighted sum over all model parameters:
  - $P(y|\mathbf{x}^*, \mathbf{y}, \mathbf{X}) = \int P(y|\mathbf{x}^*, \boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) d\boldsymbol{\theta}$

- The prediction of the MAP model is only the prediction made by the single most likely model $\boldsymbol{\theta}_{MAP}$.
  - Predictive distribution $P(y|\mathbf{x}^*, \boldsymbol{\theta}_{MAP})$.
  - Most likely prediction $f_{\theta_{MAP}}(\mathbf{x}^*) = \arg\max_y P(y|\mathbf{x}^*, \boldsymbol{\theta}_{MAP})$.

- The MAP model $\boldsymbol{\theta}_{MAP}$ is an approximaton of this weighted sum by its element with the highest weight.

Bayes-Optimal Prediction

- Bayes-optimal decision is made by a weighted sum over all model parameters:
  - $P(y|\mathbf{x}^*, \mathbf{y}, \mathbf{X}) = \int P(y|\mathbf{x}^*, \boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) d\boldsymbol{\theta}$

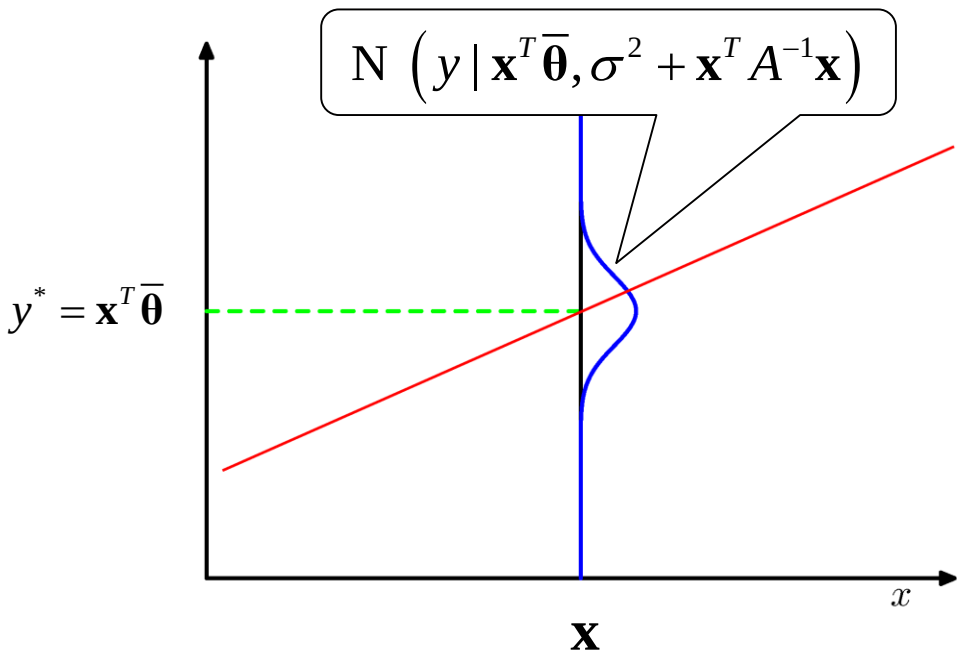
- Integration over the space of all model parameters is not generally possible.
- In some cases, there is a closed-form solution.
- In other cases, approximate numerical integration may be possible.

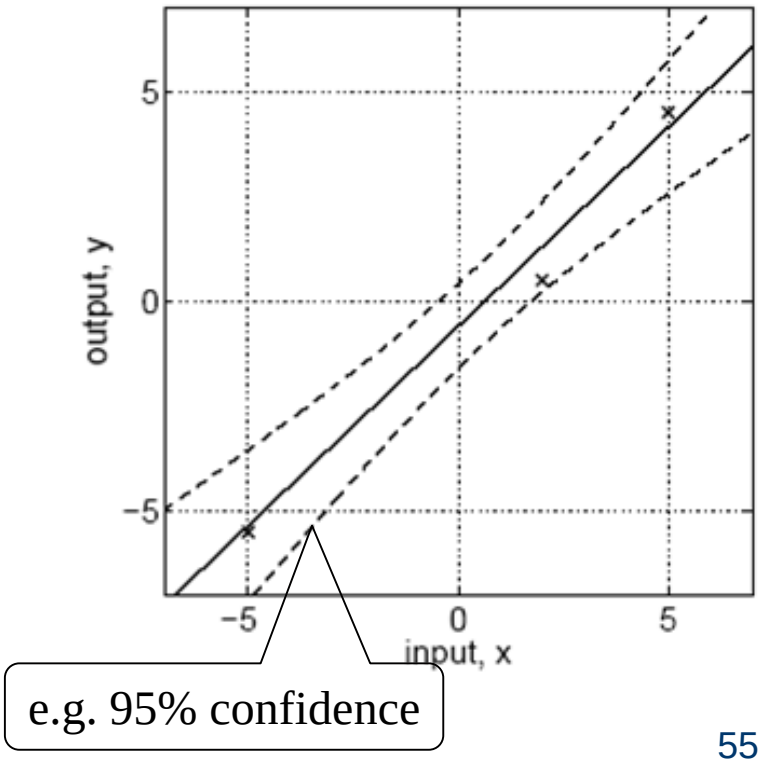
Predictive Distribution for Linear Regression

- Predictive distribution for linear regression
  - $P(y|\mathbf{x}^*, \mathbf{y}, \mathbf{X}) = \int P(y|\mathbf{x}^*, \boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) d\boldsymbol{\theta}$
    $= \int N(y|\mathbf{x}^*, \boldsymbol{\theta}) N(\boldsymbol{\theta}|\overline{\boldsymbol{\theta}}, \mathbf{A}^{-1}) d\boldsymbol{\theta}$
    $= \boldsymbol{N}(y|\overline{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{x}^*, \sigma^2 + \mathbf{x}^{*\mathrm{T}}\mathbf{A}^{-1}\mathbf{x}^*)$
  - With $\overline{\boldsymbol{\theta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1} \mathbf{X}^{\mathrm{T}}\boldsymbol{y}$
  - And $\mathbf{A}^{-1} = \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \sigma_p^{-2}\mathbf{I}$.
- Bayes-optimal prediction:
  - $y^* = \arg\max_y P(y|\mathbf{x}^*, \mathbf{y}, \mathbf{X}) = \overline{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{x}^*$

Predictive Distribution: Confidence Band

- Bayesian regression not only yields prediction $y^* = \mathbf{x}^T \overline{\boldsymbol{\theta}}$, but a distribution over y und therefore a confidence band.





Overview

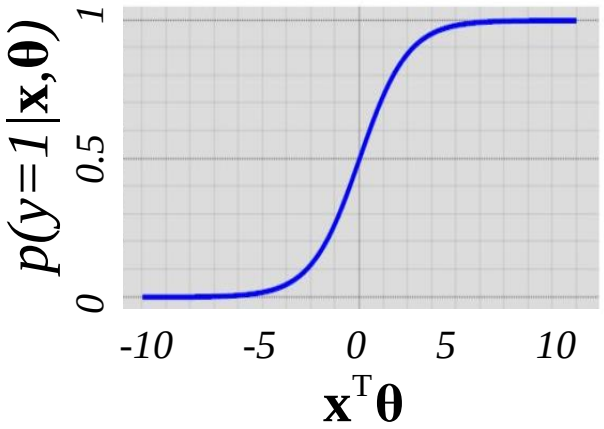
- Basic concepts of Bayesian learning
- Linear regression:
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Posterior distribution over models,
  - Bayesian prediction, predictive distribution,
- Linear classification: logistic regression.
  - Predictive distribution,
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Bayesian Prediction.
- Naive Bayesian classifier.
- Nonlinear models: Gaussian processes.

Linear Classification

- Training data:
  - $\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}$ 
  - $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ ☺ ☹
- Decision function:
  - $f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}$
- Predictive distribution
  - $P(y|\mathbf{x}, \boldsymbol{\theta}) = \sigma(\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta})$
- Linear classifier:
  - $y_{\boldsymbol{\theta}}(\mathbf{x}) = \arg\max_{y} P(y|\mathbf{x}, \boldsymbol{\theta})$
  - $y_{\boldsymbol{\theta}}$:  $\mapsto$ ☹

Linear Classification: Predictive Distribution

- For binary classification, $y \in \{-1, +1\}$
- Predictive distribution given parameters $\boldsymbol{\theta}$ of linear model:
  - $P(y = +1|\mathbf{x}, \boldsymbol{\theta}) = \sigma\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}\right) = \frac{1}{1+e^{-\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}}}$
  - $P(y = -1|\mathbf{x}, \boldsymbol{\theta}) = 1 - P(y = +1|\mathbf{x}, \boldsymbol{\theta})$
- Sigmoid function maps $[-\infty, +\infty] \rightarrow [0,1]$.



Logistic Regression

- For binary classification, $y \in \{-1, +1\}$
- Predictive distribution given parameters $\boldsymbol{\theta}$ of linear model:
  - $P(y = +1|\mathbf{x}, \boldsymbol{\theta}) = \sigma\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}\right) = \frac{1}{1+\mathrm{e}^{-\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}}}$
  - $P(y = -1|\mathbf{x}, \boldsymbol{\theta}) = 1 - \frac{1}{1+\mathrm{e}^{-\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}}} = \frac{1}{1+\mathrm{e}^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}}}$
- Written jointly for both classes:
  - $P(y|\mathbf{x}, \boldsymbol{\theta}) = \sigma\left(y\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}\right) = \frac{1}{1+\mathrm{e}^{-y\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}}}$
- Classification function:
  - $y_{\boldsymbol{\theta}}(\mathbf{x}) = \arg\max_{y} P(y|\mathbf{x}, \boldsymbol{\theta})$
- Called “logistic regression” even though it is a classification model.

Logistic Regression

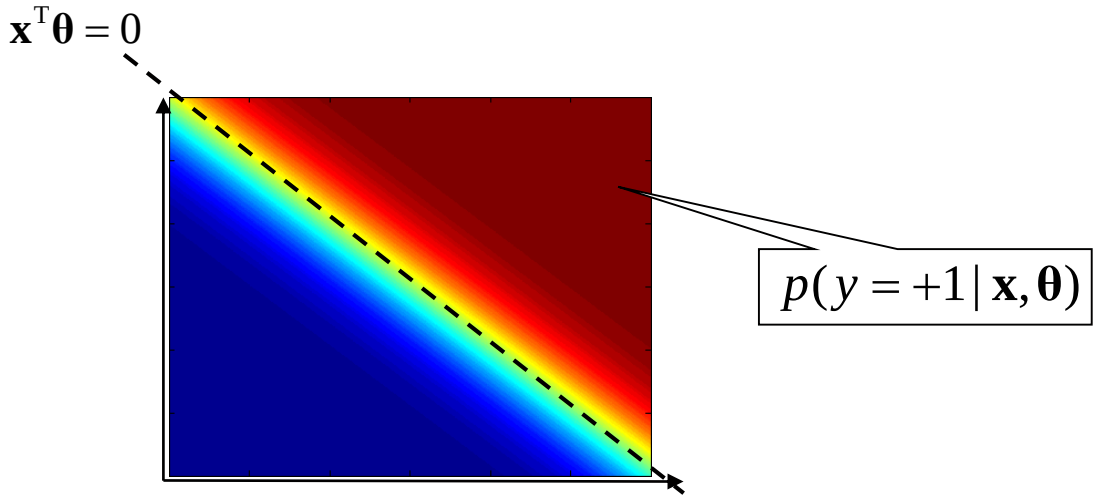
- Decision boundary: $P(y = +1|\mathbf{x}, \boldsymbol{\theta}) = P(y = -1|\mathbf{x}, \boldsymbol{\theta}) = 0.5.$

$$0.5 = \sigma\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}\right) = \frac{1}{1 + \mathrm{e}^{-\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}}}$$

$$\Leftrightarrow 1 = \mathrm{e}^{-\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}}$$

$$\Leftrightarrow 0 = \mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}$$

- Decision boundary is a hyperplane in input space.



Logistic Regression

- For multi-class classification: $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_k \end{pmatrix}$
- Generalize sigmoid function to softmax function:
  - $P(y|\mathbf{x}, \boldsymbol{\theta}) = \frac{e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}_y}}{\sum_{y'} e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}_{y'}}}$

Normalizer ensures that

$$\sum_y P(y|\mathbf{x}, \boldsymbol{\theta}) = 1$$

- Classification function:
  - $y_{\boldsymbol{\theta}}(\mathbf{x}) = \arg\max_y P(y|\mathbf{x}, \boldsymbol{\theta})$
- Called multi-class “logistic regression” even though it is a classification model.

Overview

- Basic concepts of Bayesian learning
- Linear regression:
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Posterior distribution over models,
  - Bayesian prediction, predictive distribution,
- Linear classification: logistic regression.
  - Predictive distribution,
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Bayesian Prediction.
- Naive Bayesian classifier.
- Nonlinear models: Gaussian processes.

Logistic Regression: ML Model

- Maximum-likelihood model:
  - $\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$

$$= \operatorname{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^{n} \frac{1}{1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}}$$

$$= \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} -\log \frac{1}{1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}}$$

$$= \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}\right)$$

- Equivalent to minimization of cross-entropy loss $\ell\left(\sigma\left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}\right), y_i\right) = -\log(\sigma\left(y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}\right))$
- No analytic solution; numeric optimization, for instance, using (stochastic) gradient descent.

Logistic Regression: ML Model

- Maximum-likelihood model:
  - $\boldsymbol{\theta}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}\right)$
- Gradient:
  - $\frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}\right)$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial\left(1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}\right)} \log\left(1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}\right) \frac{\partial}{\partial\left(-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}\right)} \left(1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}\right) \frac{\partial}{\partial \boldsymbol{\theta}} \left(-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}\right)$$

$$= \sum_{i=1}^{n} \frac{1}{1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}} \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}} \left(-y_i \mathbf{x}_i\right) = \sum_{i=1}^{n} -y_i \mathbf{x}_i \frac{\mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}}{1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}}$$

$$= \sum_{i=1}^{n} -y_i \mathbf{x}_i \frac{1}{1 + \mathrm{e}^{y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}}$$

$$= \sum_{i=1}^{n} -y_i \mathbf{x}_i \left(1 - \sigma\left(y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}\right)\right)$$

Overview

- Basic concepts of Bayesian learning
- Linear regression:
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Posterior distribution over models,
  - Bayesian prediction, predictive distribution,
- Linear classification: logistic regression.
  - Predictive distribution,
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Bayesian Prediction.
- Naive Bayesian classifier.
- Nonlinear models: Gaussian processes.

Logistic Regression: MAP Model

- Maximum-a-posteriori model with prior $P(\boldsymbol{\theta}) = N(\boldsymbol{\theta}|\mathbf{0}, \sigma^2\mathbf{I})$:
  - $\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) P(\boldsymbol{\theta})$

$$= \operatorname{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^{n} \frac{1}{1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}} N(\boldsymbol{\theta}|\mathbf{0}, \sigma^2\mathbf{I})$$

$$= \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} -\log \frac{1}{1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}} - \log N(\boldsymbol{\theta}|\mathbf{0}, \sigma^2\mathbf{I})$$

$$= \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\theta}}\right) + \frac{1}{2\sigma^2} \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\theta}$$

Likelihood ≈ loss function

Prior ≈ regularizer

- No analytic solution; numeric optimization, for instance, using (stochastic) gradient descent.

Logistic Regression: MAP Model

- Maximum-a-posteriori model with prior $P(\boldsymbol{\theta}) = N(\boldsymbol{\theta}|\mathbf{0}, \sigma^2\mathbf{I})$:
  - $\boldsymbol{\theta}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}}\right) + \frac{1}{2\sigma_p^2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}$
- Gradient:
  - $\frac{\partial}{\partial\boldsymbol{\theta}}\left(\sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}}\right) + \frac{1}{2\sigma_p^2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}\right)$

$$= \sum_{i=1..n} -y_i \mathbf{x}_i \left(1 - \sigma\left(y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}\right)\right) + \frac{1}{\sigma_p^2}\boldsymbol{\theta}$$

Overview

- Basic concepts of Bayesian learning
- Linear regression:
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Posterior distribution over models,
  - Bayesian prediction, predictive distribution,
- Linear classification: logistic regression.
  - Predictive distribution,
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Bayesian Prediction.
- Naive Bayesian classifier.
- Nonlinear models: Gaussian processes.

Bayes-Optimal Prediction for Classification

- Predictive distribution given the data
  - $P(y|\mathbf{x}^*, \mathbf{y}, \mathbf{X}) = \int P(y|\boldsymbol{\theta}, \mathbf{x}^*) P(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) \mathrm{d}\boldsymbol{\theta}$

$$= \int \frac{1}{1 + \mathrm{e}^{-y\mathbf{x}^{*\mathrm{T}}\boldsymbol{\theta}}} \prod_{i=1}^{n} \frac{1}{1 + \mathrm{e}^{-y_i \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\theta}}} N(\boldsymbol{\theta}|\mathbf{0}, \sigma^2\mathbf{I}) d\boldsymbol{\theta}$$

- No closed-form solution for logistic regression.
- Possible to approximate by sampling from the posterior.
- Standard approximation: use only MAP model instead of integrating over model space.

Overview

- Basic concepts of Bayesian learning
- Linear regression:
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Posterior distribution over models,
  - Bayesian prediction, predictive distribution,
- Linear classification: logistic regression.
  - Predictive distribution,
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Bayesian Prediction.
- Naive Bayesian classifier.
- Nonlinear models: Gaussian processes.

Naive Bayesian Classifier

- Decision function of Naive Bayesian classifier:

$$\max_y P(y|\mathbf{x}) = \max_y P(\mathbf{x}|y)P(y)$$

$$= \max_y P(x_1|y)P(x_2|x_1, y), \dots, P(x_m|x_1, \dots, x_{m-1}, y)$$

$$= \max_y \prod_{j=1}^{m} P\left(x_j|, x_1, \dots, x_{j-1}, y\right)P(y)$$

$$\approx \max_y \prod_{j=1}^{m} P\left(x_j|y\right)P(y)$$

Assumption: attributes are conditionally independent given the class.

$$= \max_y \prod_{j=1}^{m} \prod_{v} P\left(x_j = v|y\right)^{\left[\left[x_j=v\right]\right]} P(y)$$

Naive Bayesian Classifier

- Decision function of Naive Bayesian classifier:

$$\max_y P(y|\mathbf{x}) = \max_y P(\mathbf{x}|y)P(y)$$
$$= \max_y P(x_1|y)P(x_2|x_1, y), \dots, P(x_m|x_1, \dots, x_{m-1}, y)$$
$$= \max_y \prod_{j=1}^{m} P\left(x_j|, x_1, \dots, x_{j-1}, y\right)P(y)$$
$$\approx \max_y \prod_{j=1}^{m} \prod_v P\left(x_j = v|y\right)^{\left[[x_j=v]\right]} P(y)$$
$$= \max_y \prod_{j=1}^{m} \prod_v \theta_{x_j=v,y}^{\left[[x_j=v]\right]} \theta_y$$

These probabilities are the model parameters that have to be learned from the training data.

Naive Bayesian Classifier

- Decision function of Naive Bayesian classifier:

$$\max_y P(y|\mathbf{x}) = \max_y P(\mathbf{x}|y)P(y)$$

$$= \max_y P(x_1|y)P(x_2|x_1, y), \dots, P(x_m|x_1, \dots, x_{m-1}, y)$$

$$= \max_y \prod_{j=1}^{m} P\big(x_j|, x_1, \dots, x_{j-1}, y\big)P(y)$$

$$\approx \max_y \prod_{j=1}^{m} \prod_v P\big(x_j = v|y\big)^{\left[[x_j=v]\right]} P(y)$$

$$= \max_y \prod_{j=1}^{m} \prod_v \theta_{x_j=v,y}^{\left[[x_j=v]\right]} \theta_y$$

$$= \max_y \sum_{j=1}^{m} \sum_v \log \theta_{x_j=v,y} \big[[x_j = v]\big] + \log \theta_y$$

This is the decision function of a linear classifier.

Learning in Naive Bayes

- Maximum likelihood estimate of θ_y:
  - $\widehat{\theta_y} = \frac{n_y}{n}$, where n_y is the number of instances of class y.
- Maximum likelihood estimate of $\theta_{x_j=v,y}$:
  - $\widehat{\theta_{x_j=v,y}} = \frac{n_{x_j=v,y}}{n_y}$, where $n_{x_j=v,y}$ is the number of instances with $x_j = v$ and the class is y.
- Learning amounts to counting the frequency of all attribute values per class.

Naive Bayesian Classifier

- Naive Bayes is a linear classifier and has the same restrictions as the perceptron and other linear models.
- Naive Bayes assumes that all attributes are independent of each other given the class, which is highly unrealistic.
- There is absolutely no need to make this assumption: the perceptron and other linear classifiers do not need it.
- Naive Bayes offers no advantage over the perceptron, SVM, or logistic regression, but the independence assumption is a huge disadvatage.
- There is no reason to ever use Naive Bayes.

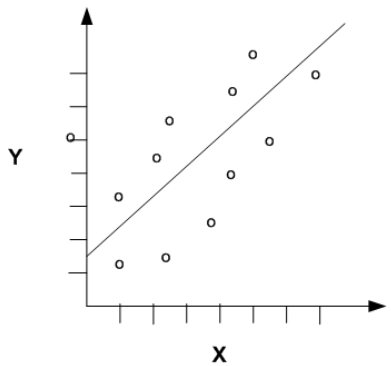
Overview

- Basic concepts of Bayesian learning
- Linear regression:
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Posterior distribution over models,
  - Bayesian prediction, predictive distribution,
- Linear classification: logistic regression.
  - Predictive distribution,
  - Maximum-likelihood model,
  - Maximum-a-posteriori model,
  - Bayesian Prediction.
- Naive Bayesian classifier.
- Nonlinear models: Gaussian processes.

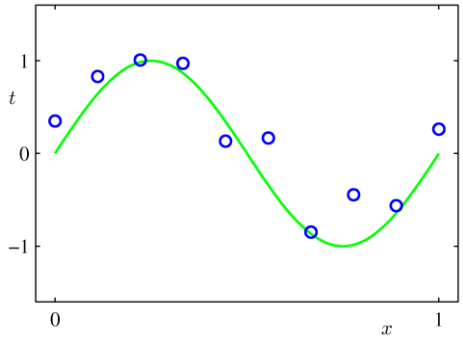
Nonlinear Regression

- Limitation of model discussed so far: only linear dependency between $\mathbf{x}$ and $f_{\boldsymbol{\theta}}(\mathbf{x})$.

Linear model



Nonlinear model



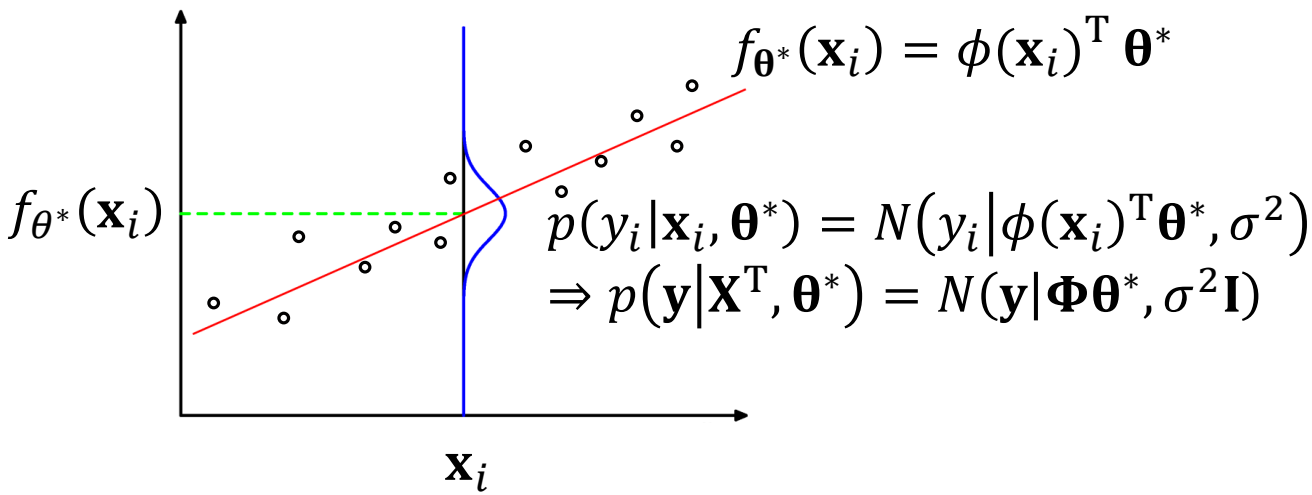
- Now: nonlinear models.

Feature Mappings and Kernels

- Use mapping ϕ to embed instances $\mathbf{x} \in X$ in higher-dimensional feature space.
- Find linear model in higher-dimensional space, corresponds to non-linear model in input space X.
- Representer theorem:
  - Model $f_{\boldsymbol{\theta}^*}(\mathbf{x}) = \boldsymbol{\theta}^{*\mathrm{T}}\phi(\mathbf{x})$
  - Has a representation $f_{\boldsymbol{\alpha}^*}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i^* \underbrace{\phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x})}_{=k(\mathbf{x}_i,\mathbf{x})}$
- Feature mapping $\phi(\mathbf{x})$ does not have to be computed; only kernel function $k(\mathbf{x}_i, \mathbf{x})$ is evaluated.
- Feature mapping $\phi(\mathbf{x})$ can therefore be high- or even infinite-dimensional.

Generalized Linear Regression (Finite-Dimensional Case)

- Assumption 1: Nature generates parameter $\boldsymbol{\theta}^*$ of a linear function $f_{\boldsymbol{\theta}^*}(\mathbf{x}) = \phi(\mathbf{x})^{\mathrm{T}}\boldsymbol{\theta}^*$ according to $p(\boldsymbol{\theta}) = N(\boldsymbol{\theta}|\mathbf{0}, \sigma_p^2\mathbf{I})$.
- Assumption 2: Inputs are $\mathbf{X}$ with feature representation $\boldsymbol{\Phi}$; line i of $\boldsymbol{\Phi}$ contains row vector $\phi(\mathbf{x}_i)^{\mathrm{T}}$. Nature generates outputs $\mathbf{y}$:
  - $y_i = f_{\boldsymbol{\theta}^*}(\mathbf{x}_i) + \epsilon_i$ with $\epsilon_i \sim N(\epsilon|0, \sigma^2)$.
  - $p(y_i|\mathbf{x}_i, \boldsymbol{\theta}^*) = N(y_i|\phi(\mathbf{x}_i)^{\mathrm{T}}\boldsymbol{\theta}^*, \sigma^2)$



Generalized Linear Regression

- Generalized linear model:
  - $f_{\boldsymbol{\theta}}(\mathbf{x}) = \phi(\mathbf{x})^{\mathrm{T}}\boldsymbol{\theta}$
  - $\mathbf{y} = \boldsymbol{\Phi}\boldsymbol{\theta}$
- Parameter $\boldsymbol{\theta}$ governed by normal distribution $N(\boldsymbol{\theta}|\mathbf{0}, \sigma_{\mathbf{p}}^2\mathbf{I})$.
- Therefore output vector $\mathbf{y}$ is also normally distributed.
  - Mean value $\mathrm{E}[\mathbf{y}] = E[\boldsymbol{\Phi}\boldsymbol{\theta}] = \boldsymbol{\Phi}E[\boldsymbol{\theta}] = \mathbf{0}$.
  - Covariance $\mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}] = \boldsymbol{\Phi}E[\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}]\boldsymbol{\Phi}^{\mathrm{T}} = \sigma_p^2\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} = \sigma_p^2\mathbf{K}$.
  - With $K_{ij} = \phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j)$.

- Remember: $\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} = \begin{bmatrix} - & \phi(\mathbf{x}_1) & - \\ & \dots & \\ - & \phi(\mathbf{x}_n) & - \end{bmatrix} \begin{bmatrix} | & & | \\ \phi(\mathbf{x}_1) & \vdots & \phi(\mathbf{x}_n) \\ | & & | \end{bmatrix} = \mathbf{K}$

Generalized Linear Regression (General Case)

- Data generation assumptions:
  - Given inputs $\mathbf{X}$, nature generates target values $\bar{\mathbf{y}} \sim N(\bar{\mathbf{y}}|0, \sigma_p^2 \mathbf{K})$.
  - Then, nature generates observations $y_i = \bar{\mathbf{y}}_i + \epsilon_i$ with noise $\epsilon_i \sim N(\epsilon|0, \sigma^2)$.

- Bayesdian inference: determine predictive distribution $P(\mathbf{y}^*|\mathbf{x}^*, \mathbf{y}, \mathbf{X})$ for new test instance.

Reminder: Linear Regression

- Bayes-optimal prediction:
  - $y^* = \arg\max_y P(y|\mathbf{x}^*, \mathbf{y}, \mathbf{X}) = \overline{\boldsymbol{\theta}}^{\mathrm{T}}\mathbf{x}^*$
  - With $\overline{\boldsymbol{\theta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$.

- Number of parameters θ_j = number of attributes in $\mathbf{x}$.

Generalized Linear Regression

- Mean value of predictive distribution $P(\mathbf{y}^*|\mathbf{x}^*, \mathbf{y}, \mathbf{X})$ has the form
  - $y^* = \sum_{i=1}^{n} \bar{\alpha}_i k(\mathbf{x}_i, \mathbf{x}^*)$
  - With $\bar{\boldsymbol{\alpha}} = \left(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1} \boldsymbol{y} = \left(\mathbf{K} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1} \boldsymbol{y}$.

- Number of parameters α_i = number of training instances.

Example nonlinear regression

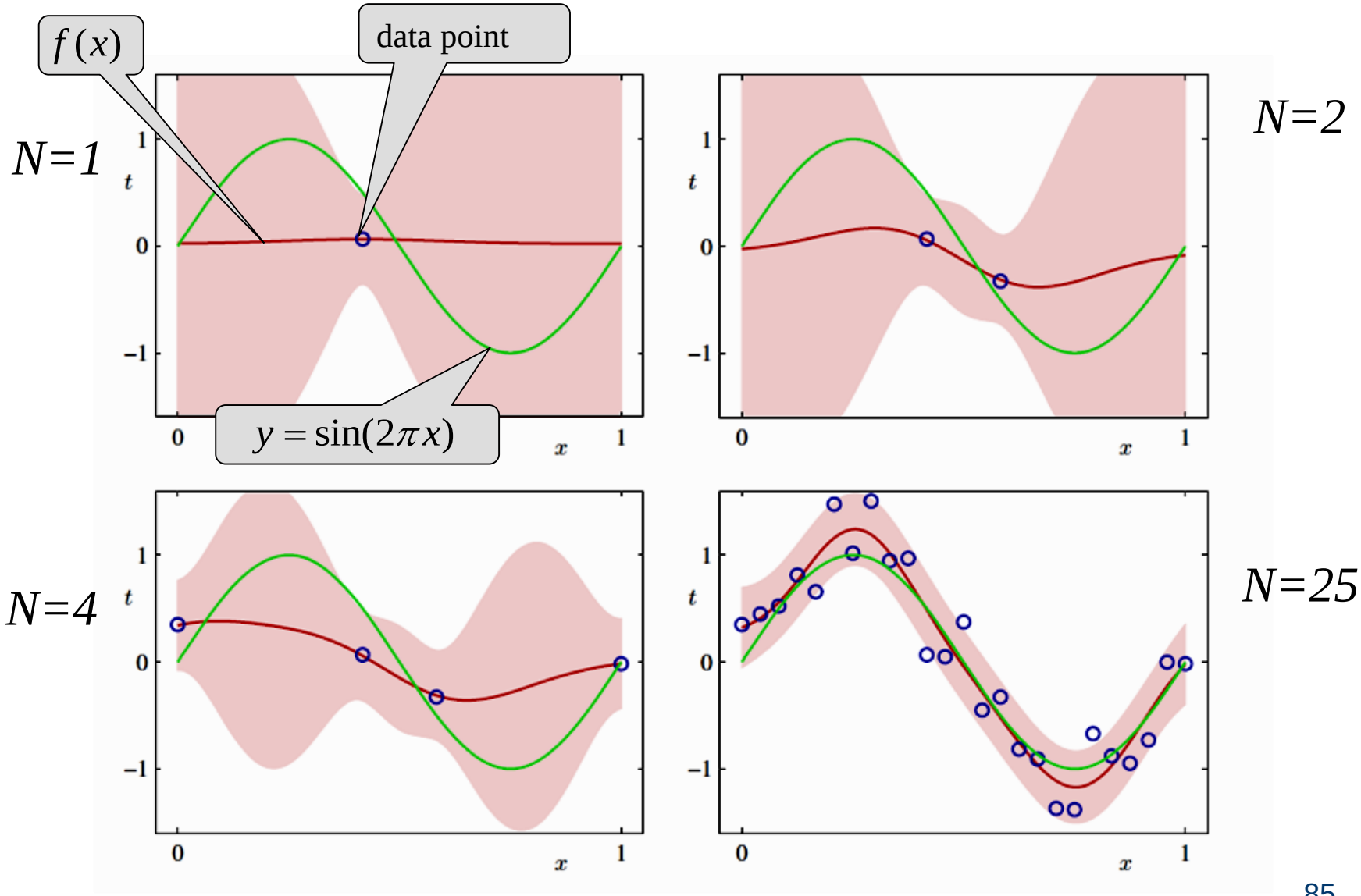
- Example for nonlinear regression
  - Generating nonlinear data by

$$y = \sin(2\pi x) + \varepsilon \qquad \varepsilon \sim \mathrm{N}\,(\varepsilon \mid 0, \sigma^2), \quad x \in [0,1]$$

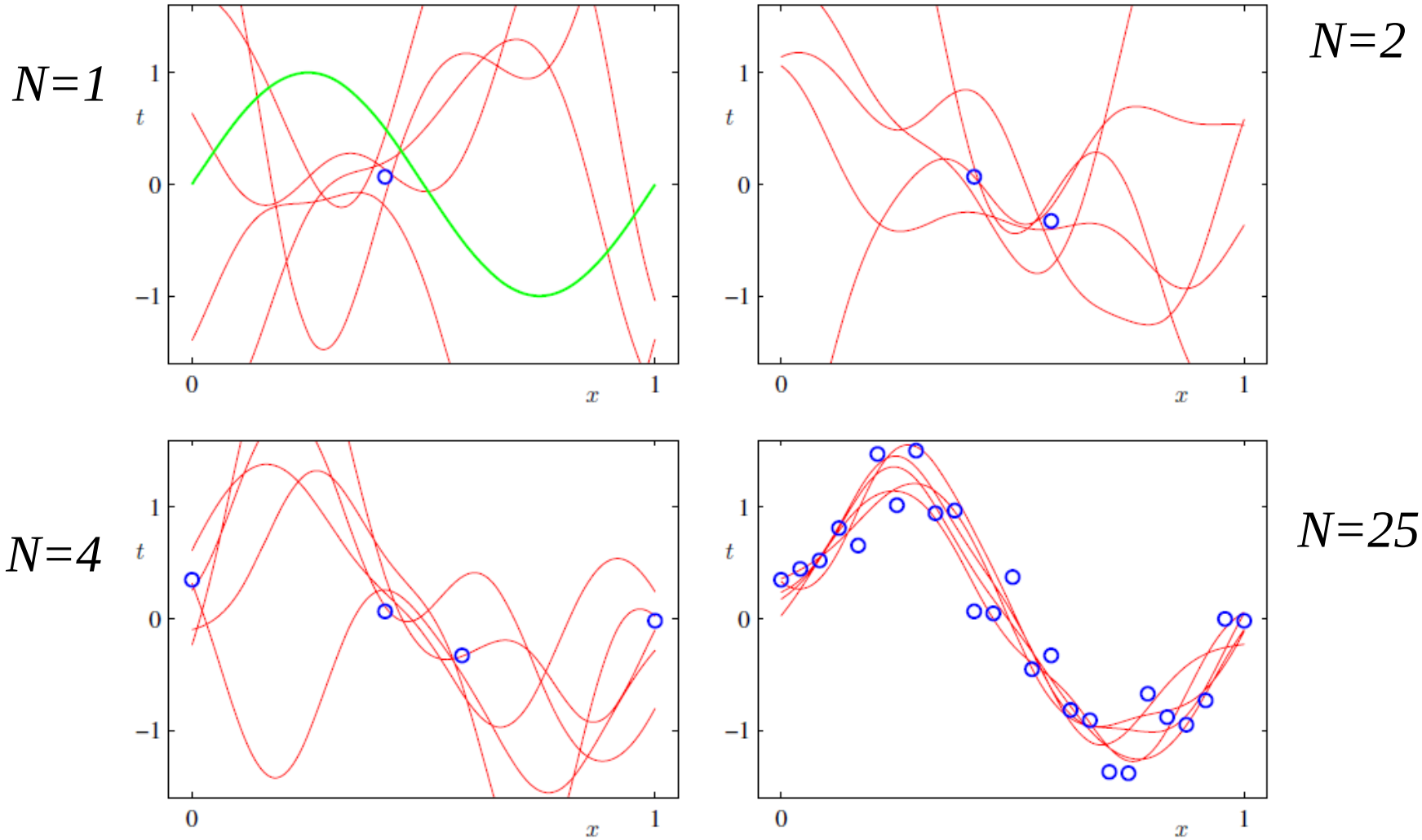
- Nonlinear kernel
  - $k(x, x') = \exp(-\theta |x - x'|)$

- How does the predictive distribution and the posterior over models look like?

Predictive distribution



Models sampled from posterior



Summary

- Linear regression:
  - Maximum-likelihood model $\arg\max_{\boldsymbol{\theta}} P(\mathbf{y}\,|\mathbf{X}, \boldsymbol{\theta})$,
  - Maximum-a-posteriori model $\arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$,
  - Posterior distribution over models $P(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$,
  - Bayesian prediction, predictive distribution $\arg\max_{y} P(\mathbf{y}^*|\mathbf{x}^*, \mathbf{y}, \mathbf{X})$.
- Linear classification (logistic regression):
  - Predictive distribution $P(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\theta})$,
  - Maximum-likelihood model $\arg\max_{\boldsymbol{\theta}} P(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$,
  - Maximum-a-posteriori model $\arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})$,
  - Bayesian Prediction $\arg\max_{y} P(\mathbf{y}|\mathbf{x}^*, \mathbf{y}, \mathbf{X})$.
- Nonlinear models: Gaussian processes.