SF1935 – Probability Theory and Statistics with **Application to Machine Learning**

Lecture 2: Linear models and a probabilistic perspective

Pawel Herman

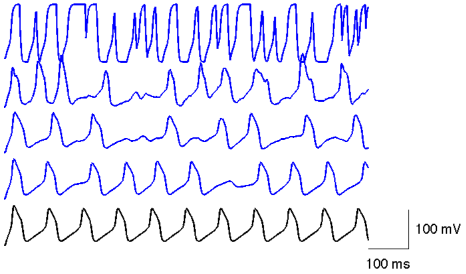
Computational Science and Technology (CST)

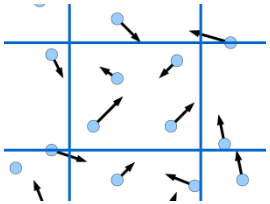
KTH Royal Institute of Technology

Introduction

Probabilistic approach

- Ubiquitous nature of uncertainty
  - imprecision, noise in data,
  - errors , missing information/data
  - gaps in knowledge, simplified description



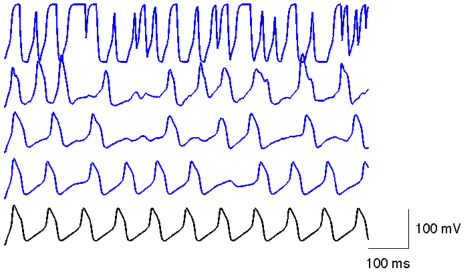


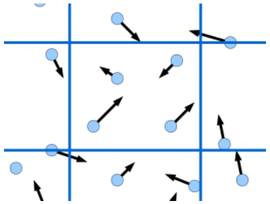
Introduction

Probabilistic approach

➢ Ubiquitous nature of uncertainty

- imprecision, noise in data,
- errors , missing information/data
- gaps in knowledge, simplified description



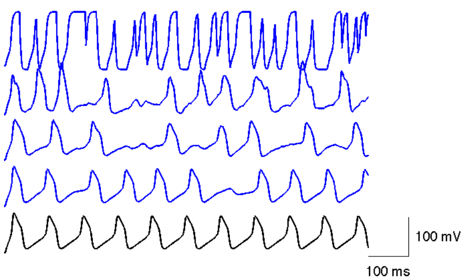


"The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which of times they are unable to account." (Laplace)

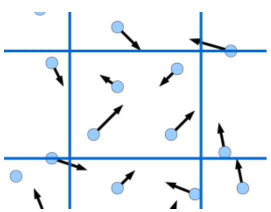
Introduction

Probabilistic approach

- ➢ Ubiquitous nature of uncertainty
  - imprecision, noise in data
  - errors , missing information/data
  - gaps in knowledge, simplified description



- ➢ Probability theory provides a framework for modelling, reasoning etc. under uncertainty
  - unified, universal, intuitive, interpretable
  - beyond randomness, it is about uncertainty!
  - p. distributions as carriers or information (Jaynes, 2003)



Probabilistic perspective in ML

- Statistical ML: constructing stochastic models
  - fully probabilistic description and inference
  - theoretical assumptions, mathematical tractability, rigour
  - parameters estimated from observed data (learning)
  - interpretability and extra insights
  - machinery to propagate and account for uncertainty effects

Probabilistic perspective in ML

- Statistical ML: constructing stochastic models
  - fully probabilistic description and inference
  - theoretical assumptions, mathematical tractability, rigour
  - parameters estimated from observed data (learning)
  - interpretability and extra insights
  - machinery to propagate and account for uncertainty effects

BUT: *can be very hard for large-scale problems* and

difficult to derive solutions in a closed form

Probabilistic perspective in ML

- Statistical ML: constructing stochastic models
  - fully probabilistic description and inference
  - t igour
  - p
  - i
  - n effects

“ (statistical ML) provides the basis for learning algorithms that directly manipulate probabilities, as well as a framework for analyzing the operation of other algorithms that do not explicitly manipulate probabilities.” Mitchell

Probabilistic perspective in ML

- Philosophy of Bayesian approach
  - Uncertainty is ubiquitous – describe all model components with probabilistic objects (*distributions, not point estimates*)

Probabilistic perspective in ML

- Philosophy of Bayesian approach
  - Uncertainty is ubiquitous – describe all model components with probabilistic objects (*distributions, not point estimates*)
  - Apply Bayesian machinery to propagate uncertainty
    - product and sum probability rules, Bayesian theorem

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

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Probabilistic perspective in ML

- Philosophy of Bayesian approach
  - Uncertainty is ubiquitous – describe all model components with probabilistic objects (*distributions, not point estimates*)
  - Apply Bayesian machinery to propagate uncertainty
    - product and sum probability rules, Bayesian theorem
    - the power of marginalisation

$$p(x_1, x_2 .., x_{n-1}) = \int p(x_1, x_2, .., x_n) dx_n$$

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

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Probabilistic perspective in ML

- Philosophy of Bayesian approach
  - Uncertainty is ubiquitous – describe all model components with probabilistic objects (*distributions, not point estimates*)
  - Apply Bayesian machinery to propagate uncertainty
  - Combine uncertain knowledge with data to reduce uncertainty (based on evidence from observations)
    - batch or sequence where posterior is iteratively updated

Probabilistic perspective in ML

- Philosophy of Bayesian approach
  - Uncertainty is ubiquitous – describe all model components with probabilistic objects (*distributions, not point estimates*)
  - Apply Bayesian machinery to propagate uncertainty
  - Combine uncertain knowledge with data to reduce uncertainty (based on evidence from observations)
  - Two levels of inference:

    parameter estimation and model selection (see Lecture 3)

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)

I. The ratio of outcomes in repeated trials.

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)

II. The data is fixed, models have probabilities

I. The ratio of outcomes in repeated trials.

II. There is a true model and the data is a random realisation.

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)
II. The data is fixed, models have probabilities
III. There does not have to be an experiment for declaring probability.

I. The ratio of outcomes in repeated trials.
II. There is a true model and the data is a random realisation.
III. Parameters can only be deduced from data (likely outcome of exp.)

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)

II. The data is fixed, models have probabilities

III. There does not have to be an experiment for declaring probability.

IV. Can incorporate prior knowledge, probabilities can be updated.

I. The ratio of outcomes in repeated trials.

II. There is a true model and the data is a random realisation.

III. Parameters can only be deduced from observed data (likely outcome of exp.)

IV. Each repeated experiment starts from ignorance.

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)
II. The data is fixed, models have probabilities
III. There does not have to be an experiment for declaring probability.
IV. Can incorporate prior knowledge, probabilities can be updated.
V. Estimators are good for available data.

I. The ratio of outcomes in repeated trials.
II. There is a true model and the data is a random realisation.
III. Parameters can only be deduced from observed data (likely outcome of exp.)
IV. Each repeated experiment starts from ignorance.
V. Estimators are averaged across many trials.

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)
II. The data is fixed, models have probabilities
III. There does not have to be an experiment for declaring probability.
IV. Can incorporate prior knowledge, probabilities can be updated
V. Estimators are good for available data.
VI. Probability of a hypothesis given the data (posterior distribution).

I. The ratio of outcomes in repeated trials.
II. There is a true model and the data is a random realisation.
III. Parameters can only be deduced from observed data (likely outcome of exp.)
IV. Each repeated experiment starts from ignorance.
V. Estimators are averaged across many trials.
VI. Probability of the data given hypothesis (likelihood, sampling dist.).

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)

II. The data is fixed, models have probabilities

III. There does not have to be an experiment for declaring probability.

IV. Can incorporate prior knowledge, probabilities can be updated

V. Estimators are good for available data.

VI. Probability of a hypothesis given the data.

VII. All variables/parameters have distribution.

I. The ratio of outcomes in repeated trials.

II. There is a true model and the data is a random realisation.

III. Parameters can only be deduced from observed data (likely outcome of exp.)

IV. Each repeated experiment starts from ignorance.

V. Estimators are averaged across many trials.

VI. Probability of the data given hypothesis.

VII. Parameters are fixed unknowns that can be point estimated from repeated trials.

Back to the probabilistic mindset

Bayesian vs frequentist perspective

I. Probability is a measure of belief (plausibility)

II. The data is fixed, models have probabilities

III. There does not
declaring proba

IV. Can incorporate
probabilities can be updated

V. Estimators are good for available data.

VI. Probability of a hypothesis given the data.

VII. All variables/parameters have distribution.

I. The ratio of outcomes in repeated trials.

II. There is a true model and the data is a random

uced from
ne of exp.)

IV. Each repeated experiment starts from ignorance.

V. Estimators are averaged across many trials.

VI. Probability of the data given hypothesis.

VII. Parameters are fixed unknowns that can be point estimated from repeated trials.

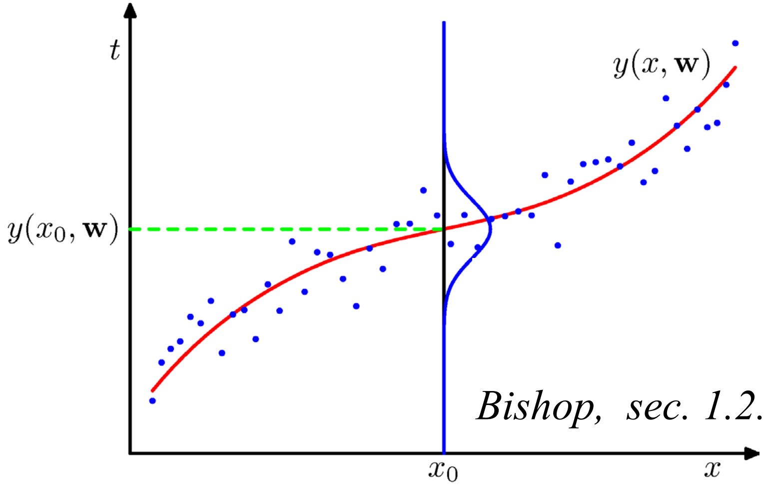
Why isn't everyone Bayesian? *(Efron, 1986)*

Learning as inference

- Learning distributions
  - curve-fitting example: $y(x, \mathbf{w}): X \to Y$ (let's assume polynomial)

$$t = y(x, \mathbf{w}) + \varepsilon$$

observations $\{(x_i, t_i): i=1,..,N\}$: $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$, $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$



Bishop, sec. 1.2.5

Learning as inference

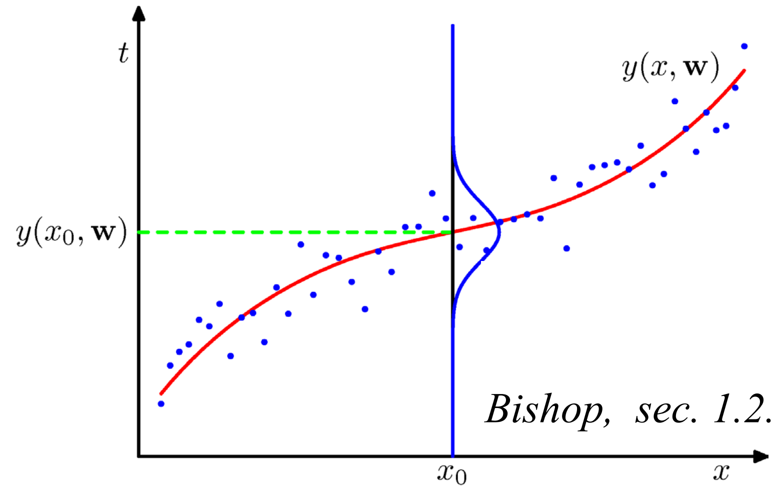
- Learning distributions
  - curve-fitting example: $y(x,\mathbf{w}) : X \rightarrow Y$ (let's assume polynomial)

$$t = y(x,\mathbf{w}) + \varepsilon$$

observations $\{(x_i,\, t_i)\colon i=1,..,N\}$: $\mathbf{x} = (x_1, \ldots, x_N)^T$, $\mathbf{t} = (t_1, \ldots, t_N)^T$

<u>Remarks about notation:</u>

1) here we deal with one-dim input, x and output, t
2) parameters $\boldsymbol{w}$ still constitute a vector (e.g. could be polynomial coefficients)
3) $\mathbf{x}$ and $\mathbf{t}$, refer to the collection of all inputs and outputs, conceptually corresponding to D_x and D_t, but in the vector form, so $\mathbf{x} \rightarrow \mathbf{t}$.



Bishop, sec. 1.2.5

Learning as inference

- Learning distributions
  - curve-fitting example: $y(x,\mathbf{w}) : X \to Y$ (let's assume polynomial)

$$t = y(x,\mathbf{w}) + \varepsilon$$

observations $\{(x_i, t_i) \colon i=1,..,N\}$: $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$, $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$

  - probabilistic framework

$p(\mathbf{w})$ – a prior probability distribution

$p(\mathcal{D} \mid \mathbf{w})$ – the likelihood function for $\mathcal{D} = \mathbf{t}$
(not a probability distrib. over $\mathbf{w}$, just a conditional probability)

$p(\mathbf{w} \mid \mathcal{D})$ – a posterior probability

Learning as inference

- Learning distributions
  - curve-fitting example: $y(x, \mathbf{w}) : X \to Y$ (let's assume polynomial)

observatio $_N)^{\mathrm{T}}$

  - probab

$p(\mathbf{w}$

Bayes' theorem

$$\color{red}{p(\mathbf{w} \mid \mathcal{D})} = \frac{\color{blue}{p(\mathcal{D} \mid \mathbf{w})}\, \color{green}{p(\mathbf{w})}}{p(\mathcal{D})}$$

$$\color{red}{\text{posterior}} \propto \color{blue}{\text{likelihood}} \times \color{green}{\text{prior}}$$

$p(\mathcal{D} \mid \mathbf{w})$ – the likelihood function for $\mathcal{D} = \mathbf{t}$
(not a probability distrib. over $\mathbf{w}$, just a conditional probability)

$p(\mathbf{w} \mid \mathcal{D})$ – a posterior probability

Learning as inference

- Learning distributions

  - curve-fitting example: $y(x,\mathbf{w}) : X \to Y$ (let's assume polynomial)

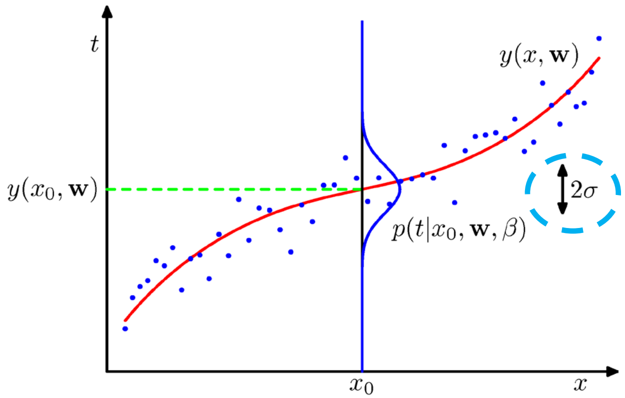
$$t = y(x,\mathbf{w}) + \varepsilon$$

observations $\{(x_i, t_i): i=1,..,N\}$: $\mathbf{x} = (x_1, \ldots, x_N)^T$, $\mathbf{t} = (t_1, \ldots, t_N)^T$

Uncertainty (*noise*) in target data:

$$p(t \mid x, \mathbf{w}, \beta) = \mathcal{N}(t \mid y(x,\mathbf{w}), \beta^{-1})$$

precision



Bishop, sec. 1.2.5

Learning as inference – estimate parameters

- Learning distributions
  - curve-fitting example: $y(x,\mathbf{w}): X \to Y$ (let's assume polynomial)

$$t = y(x,\mathbf{w}) + \varepsilon$$

observations $\{(x_i, t_i): i=1,..,N\}$: $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$, $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$

SOLUTION 1

We want to find parameters to be able to use predictive distribution, $p(t|x)$, and infer the target:

$$\mathrm{E}[t \mid x] = \int t\, p(t \mid x, \mathbf{w}_{\mathrm{opt}}, \beta_{\mathrm{opt}})\,\mathrm{d}t$$

Learning as inference – max likelihood

So, we follow the max likelihood approach

ML function for t (i.i.d.) under Gaussian noise $p(t \mid x, \mathbf{w}, \beta) = \mathcal{N}(t \mid y(x, \mathbf{w}), \beta^{-1})$

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid y_n(x_n, \mathbf{w}), \beta^{-1})$$

Learning as inference – max likelihood

So, we follow the max likelihood approach

ML function for t (i.i.d.) under Gaussian noise $p(t \mid x, \mathbf{w}, \beta) = \mathcal{N}(t \mid y(x, \mathbf{w}), \beta^{-1})$

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid y_n(x_n, \mathbf{w}), \beta^{-1})$$

The log-likelihood:

$$\boxed{\ln p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)}$$

Learning as inference – max likelihood

So, we follow the max likelihood approach

ML function for t (i.i.d.) under Gaussian noise $p(t \mid x, \mathbf{w}, \beta) = \mathcal{N}(t \mid y(x, \mathbf{w}), \beta^{-1})$

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid y_n(x_n, \mathbf{w}), \beta^{-1})$$

The log-likelihood:

$$\boxed{\ln p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)}$$

Still, input x is one-dim and $\mathbf{x}$ & $\mathbf{t}$ refer to the collection of all data points.

Learning as inference – max likelihood

Maximum likelihood (ML) estimate:

Maximise $$\ln p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

Learning as inference – max likelihood

Maximum likelihood (ML) estimate:

Maximise $$\ln p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

$$\mathbf{w}_{\mathrm{ML}} = \arg\min_{\mathbf{w}} \left\{ \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \right\}$$

the sum-of-squares error function

under the assumption of Gaussian noise : $p(t \mid x, \mathbf{w}, \beta) = \mathcal{N}(t \mid y(x, \mathbf{w}), \beta^{-1})$

Learning as inference – max likelihood

Maximum likelihood (ML) estimate:

Maximise $\ln p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$

$$\mathbf{w}_{\mathrm{ML}} = \arg\min_{\mathbf{w}} \left\{ \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \right\}$$

$$\beta_{\mathrm{ML}}^{-1} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2$$

Learning as inference – max likelihood

Maximum likelihood (ML) estimate:

Maximise $$\ln p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{ y(x_n, \mathbf{w}) - t_n \}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

$$\mathbf{w}_{\mathrm{ML}} = \arg\min_{\mathbf{w}} \left\{ \sum_{n=1}^{N} \{ y(x_n, \mathbf{w}) - t_n \}^2 \right\}$$

$$\beta_{\mathrm{ML}}^{-1} = \frac{1}{N} \sum_{n=1}^{N} \{ y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n \}^2$$

$$\boxed{t_{out} = \mathrm{E}[t \mid x_{in}] = \int t \, p(t \mid x_{in}, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) \, \mathrm{d}t}$$

Learning as inference – posterior over parameters

- Learning distributions
  - curve-fitting example: $y(x,\mathbf{w}): X \rightarrow Y$ (let's assume polynomial)

$$t = y(x,\mathbf{w}) + \varepsilon$$

SOLUTION 2

Parameters of the model could also be random variables with a prior distribution, $p(\mathbf{w}\mid\alpha) = \mathcal{N}(\mathbf{w}\mid\mathbf{0},\alpha^{-1}\mathbf{I})$

Now, we must **maximise the posterior** $p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta)$, i.e. find the most probable $\mathbf{w}$ given data:

$$\max\, p(\mathbf{w}\mid\mathbf{x},\mathbf{t},\alpha,\beta) \propto p(\mathbf{t}\mid\mathbf{x},\mathbf{w},\beta)\; p(\mathbf{w}\mid\alpha)$$

Learning as inference – max posterior

- Learning distributions
  - curve-fitting example: $y(x, \mathbf{w}) : X \rightarrow Y$ (let's assume polynomial)

$$t = y(x, \mathbf{w}) + \varepsilon$$

SOLUTION 2

Parameters of the model could also be random variables with a prior distribution, $p(\mathbf{w} \mid \alpha) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I})$

Now, we must **maximise the posterior** $p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta)$, i.e. find the most probable $\mathbf{w}$ given data:

$$\max\, p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta)\; p(\mathbf{w} \mid \alpha)$$

Again, $\boldsymbol{x}$ & $\boldsymbol{t}$ refer to the data collection, not individual input or outputs

Learning as inference – max posterior

Maximum posterior (MAP) estimate:

Maximise $\ln p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta)$

$$\boxed{\mathbf{w}_{\mathrm{MAP}} = \arg\min_{\mathbf{w}} \left\{ \frac{\beta}{2} \sum_{n=1}^{N} \{ y(x_n, \mathbf{w}) - t_n \}^2 + \frac{\alpha}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w} \right\}}$$

Learning as inference – recap

Maximum posterior (MAP) estimate:

Maximise $\ln p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta)$

$$\boxed{\mathbf{w}_{\mathrm{MAP}} = \arg\min_{\mathbf{w}} \left\{ \frac{\beta}{2} \sum_{n=1}^{N} \{ y(x_n, \mathbf{w}) - t_n \}^2 + \frac{\alpha}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w} \right\}}$$

BUT: Solutions 1 (ML) and 2 (MAP) give point estimates of $\mathbf{w}$

Learning as inference – Bayesian view

Instead of estimating “optimal” parameters $\mathbf{w}$, let’s integrate over all values of $\mathbf{w}$ (let’s make use of the distribution)

Marginalisation:

$$\underbrace{p(t \mid x, \mathbf{x}, \mathbf{t}, \alpha, \beta)}_{\text{predictive}} = \int \underbrace{p(t \mid x, \mathbf{w}, \beta)}_{\text{“noise” model}} \; \underbrace{p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta)}_{\text{posterior}} \, \mathrm{d}\mathbf{w}$$

We assume we know what α and β are.

Learning as inference – Bayesian view

Instead of estimating "optimal" parameters $\mathbf{w}$, let's integrate over all values of $\mathbf{w}$ for our linear model $y = \sum_{i=0}^{M} w_i \phi_i(x)$

For a one-dim polynomial model: $\phi_i(x) = x^i$ (order M-1)

Learning as inference – Bayesian view

Instead of estimating "optimal" parameters $\mathbf{w}$, let's integrate over all values of $\mathbf{w}$ for our linear model $y = \sum_{i=0}^{M} w_i \phi_i(x)$

For a one-dim polynomial model: $\phi_i(x) = x^i$

$$p(t \mid x, \mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}\left(t \mid m(x), s^2(x)\right) \quad \begin{cases} m(x) = \beta\, \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n)\, t_n \\ s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S}\, \boldsymbol{\phi}(x) \end{cases}$$

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^{\mathrm{T}}$$

$$\boldsymbol{\phi}(x) = (\phi_0(x), \ldots, \phi_M(x))$$

Learning as inference – summary

In summary:

To use a predictive distribution and infer the output t for the given input x_{in} ....

Learning as inference – summary

In summary:

To use a predictive distribution and infer the output t for the given input x_{in} ....

.....ML and MAP approaches produce point estimates of $\mathbf{w}$

$$\text{ML}: \quad \mathcal{D} \rightarrow \mathbf{w}_{\text{ML}}$$

$$\text{MAP}: \quad \mathcal{D}, p(\mathbf{w}) \rightarrow \mathbf{w}_{\text{MAP}}$$

Learning as inference – summary

In summary:

To use a predictive distribution and infer the output t for the given input x_{in} ....

.....ML and MAP approaches produce point estimates of $\mathbf{w}$

$$\text{ML}: \quad \mathcal{D} \to \mathbf{w}_{\text{ML}}$$

$$\text{MAP}: \quad \mathcal{D}, p(\mathbf{w}) \to \mathbf{w}_{\text{MAP}}$$

.....in Bayesian approach $\mathbf{w}$ is integrated over (**marginalisation**)

$$\text{Bayes:} \quad \mathcal{D}, p(\mathbf{w}) \to p(\mathbf{w} \mid \mathcal{D})$$

Learning as inference – summary

In summary:

To use a predictive distribution and infer the output t for the given input x_{in} ....

.....ML and MAP approaches produce point estimates of $\mathbf{w}$

$$\text{ML}: \quad \mathcal{D} \to \mathbf{w}_{\text{ML}}$$

$$\text{MAP}: \quad \mathcal{D}, p(\mathbf{w}) \to \mathbf{w}_{\text{MAP}}$$

$$\Downarrow$$

Use $p(t \mid x_{in}, \mathbf{w}_{\text{ML}})$ or $p(t \mid x_{in}, \mathbf{w}_{\text{MAP}})$ for prediction.

.....in Bayesian approach $\mathbf{w}$ is integrated over (**marginalisation**)

$$\text{Bayes}: \quad \mathcal{D}, p(\mathbf{w}) \to p(\mathbf{w} \mid \mathcal{D})$$

$$\Downarrow$$

Marginalise over $\mathbf{w}$:

$$p(t \mid \mathcal{D}) = \int p(t \mid x_{in}, \mathbf{w})\, p(\mathbf{w} \mid \mathcal{D})\, \mathrm{d}\mathbf{w}$$

Learning as inference – summary

In summary:

To use a predictive distribution and infer the output t for the given input x_{in} ....

.....ML and MAP approaches produce point estimates of $\mathbf{w}$

$$\text{ML}: \qquad \mathcal{D} \to \mathbf{w}_{\text{ML}}$$

$$\text{MAP}: \qquad \mathcal{D}, p(\mathbf{w}) \to \mathbf{w}_{\text{MAP}}$$

Frequentist philosophy

$p(t \mid x_{in}, \mathbf{w}_{\text{MAP}})$ for prediction.

.....in Bayesian approach $\mathbf{w}$ is integrated over (**marginalisation**)

$$\text{Bayes:} \quad \mathcal{D}, p(\mathbf{w}) \to p(\mathbf{w} \mid \mathcal{D})$$

Bayesian philosophy

$$p(t \mid \mathcal{D}) = \int p(t \mid x_{in}, \mathbf{w})\, p(\mathbf{w} \mid \mathcal{D})\, \mathrm{d}\mathbf{w}$$

Inference and decision (classification)

The ***inference*** stage of classification $\mathcal{D} \rightarrow p(C_k, \boldsymbol{x}), k = 1,..,K$

Inference and decision (classification)

The ***inference*** stage of classification $\mathcal{D} \rightarrow p(C_k, \boldsymbol{x}), k = 1,..,K$

starting from *class-conditional* densities and *priors*

Model the inputs $\boldsymbol{x}$ and outputs C

$$\left.\begin{matrix} p(\boldsymbol{x} \mid C_k) \\ p(C_k) \end{matrix}\right\} \text{ for each class } C_k \qquad p(\boldsymbol{x}, C_k)$$

$$p(C_k \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid C_k)\, p(C_k)}{p(\boldsymbol{x})}$$

$$p(\boldsymbol{x}) = \sum_k p(\boldsymbol{x} \mid C_k)\, p(C_k)$$

GENERATIVE approach

Remarks:
1) K classes
2) $\boldsymbol{x}$ – multi-dim input feature vector

Inference and decision (classification)

The ***inference*** stage of classification $\mathcal{D} \rightarrow p(C_k, \boldsymbol{x}), k = 1,..,K$

Model the inputs $\boldsymbol{x}$ and outputs C

$$\left.\begin{array}{l} p(\boldsymbol{x} \mid C_k) \\ p(C_k) \end{array}\right\} \text{ for each class } C_k \qquad p(\boldsymbol{x}, C_k)$$

$$p(C_k \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid C_k)\, p(C_k)}{p(\boldsymbol{x})}$$

$$p(\boldsymbol{x}) = \sum_k p(\boldsymbol{x} \mid C_k)\, p(C_k)$$

GENERATIVE approach

Solve first the inference problem of determining ***posteriors*** for each class **without** modelling $p(C_k, \boldsymbol{x})$

$$p(C_k \mid \boldsymbol{x})$$

DISCRIMINATIVE approach

Generative vs discriminative approach

What are the virtues of the generative approach?

- The parameters are estimated separately for each class (no need to retrain the model when new classes are added)
- Rather straightforward to fit in a Bayesian framework (but it depends on the problem, sometimes discriminative function can be easier to optimise)

Generative vs discriminative approach

What are the virtues of the generative approach?

- The parameters are estimated separately for each class (no need to retrain the model when new classes are added)
- Rather straightforward to fit in a Bayesian framework (but it depends on the problem, sometimes discriminative function can be easier to optimise)
- Easy and elegant way of handling missing or unlabelled data
- Generative model allows for….. generating data

  -> generative models can be run "backwards"

Generative vs discriminative approach

What are the virtues of the generative approach?

- The parameters are estimated separately for each class (no need to retrain the model when new classes are added)
- Rather straightforward to fit in a Bayesian framework (but it depends on the problem, sometimes discriminative function can be easier to optimise)
- Easy and elegant way of handling missing or unlabelled data
- Generative model allows for….. generating data

  -> generative models can be run "backwards"
- BUT: discriminative models tend to be more accurate (less vulnerable to assumptions)

Bias-variance: frequentist mindset

Expected value over all possible datasets D
(frequentist perspective: data is random)

$$\mathbb{E}[L] = \mathbb{E}_{\mathcal{D}}\left\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\right\}^2 + \int \mathrm{var}\left[t \mid \boldsymbol{x}\right] p(\boldsymbol{x})\, d\boldsymbol{x}$$

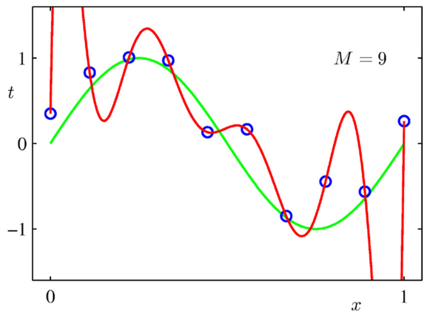
$$\left\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\right\}^2 = \left\{y(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[y(\boldsymbol{x};\mathcal{D})\right] + \mathbb{E}_{\mathcal{D}}\left[y(\boldsymbol{x};\mathcal{D})\right] - h(\boldsymbol{x})\right\}^2$$

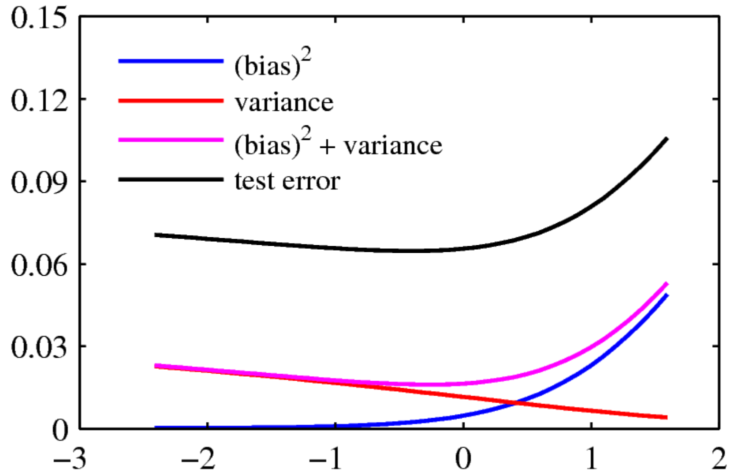
$$\mathbb{E}_{\mathcal{D}}\left[\left\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\right\}^2\right] = \underbrace{\left\{\mathbb{E}_{\mathcal{D}}\left[y(\boldsymbol{x};\mathcal{D})\right] - h(\boldsymbol{x})\right\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left\{y(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[y(\boldsymbol{x};\mathcal{D})\right]\right\}^2\right]}_{\text{variance}}$$

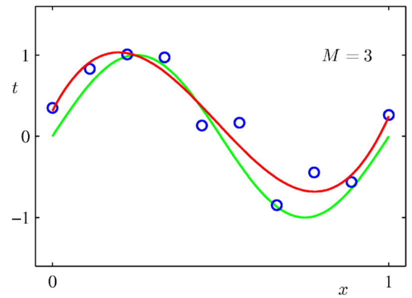
Bias-variance as a frequentist dilemma

$$\mathbb{E}_{\mathcal{D}}[L] = \mathbb{E}_{\mathcal{D}}\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\}^2 + \int \mathrm{var}[t\,|\,\boldsymbol{x}]\,p(\boldsymbol{x})\,dx$$

$$\mathbb{E}[L] = (\text{bias})^2 + \text{variance} + \text{noise}$$







increased model complexity

Bishop, sec. 3.2

Model selection: frequentist vs Bayesian

- Occam's razor – *"Accept the simplest explanation that fits the data."*
- Frequentist approach with maximum likelihood
  - bias-variance dilemma
  - need to control the model's complexity
    - regularisation
    - correction for the bias of ML estimates (AIC, BIC)
    - empirical estimate of generalisation error on a hold-out set (validation, resampling)
    - structural risk minimization (SRM) (minimise upper bound on the true risk), see also VC dimension (statistical learning theory)

Model selection: frequentist vs Bayesian

- Occam's razor – *"Accept the simplest explanation that fits the data."*
- Frequentist approach with maximum likelihood
- Bayesian approach using model evidence

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w})}{p(\mathcal{D})} \Longrightarrow p(\mathbf{w} \mid \mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D} \mid \mathbf{w}, \mathcal{M}_i)\, p(\mathbf{w} \mid \mathcal{M}_i)}{p(\mathcal{D} \mid \mathcal{M}_i)}$$

$p(\mathcal{D})$ shows where the model spreads its probability mass over the data space

Model selection: frequentist vs Bayesian

- Occam's razor – *"Accept the simplest explanation that fits the data."*
- Frequentist approach with maximum likelihood
- Bayesian approach using model evidence

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w})}{p(\mathcal{D})} \quad \Longrightarrow \quad p(\mathbf{w} \mid \mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D} \mid \mathbf{w}, \mathcal{M}_i)\, p(\mathbf{w} \mid \mathcal{M}_i)}{p(\mathcal{D} \mid \mathcal{M}_i)}$$

$p(\mathcal{D})$ shows where the model spreads its probability mass over the data space

Towards the posterior: $p(\mathcal{M}_i \mid \mathcal{D}) \propto p(\mathcal{M}_i)\, p(\mathcal{D} \mid \mathcal{M}_i)$

marginal likelihood

Model selection: frequentist vs Bayesian

- Occam's razor – *"Accept the simplest explanation that fits the data."*
- Frequentist approach with maximum likelihood
- Bayesian approach using model evidence

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w})\, p(\mathbf{w})}{p(\mathcal{D})} \Longrightarrow p(\mathbf{w} \mid \mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D} \mid \mathbf{w}, \mathcal{M}_i)\, p(\mathbf{w} \mid \mathcal{M}_i)}{p(\mathcal{D} \mid \mathcal{M}_i)}$$

calculating $p(\mathcal{D} \mid \mathcal{M}_i)$ is not so trivial:

$$p(\mathcal{D} \mid \mathcal{M}_i) = \int p(\mathcal{D} \mid \mathbf{w}, \mathcal{M}_i)\, p(\mathbf{w} \mid \mathcal{M}_i)\, d\mathbf{w}$$

"The evidence can be seen as the probability of generating the data set from a model whose parameters are sampled at random from the prior"