

RWTH AACHEN UNIVERSITY

Advanced Machine Learning Lecture 2

Linear Regression

29.10.2015

Bastian Leibe
RWTH Aachen
<http://www.vision.rwth-aachen.de/>
leibe@vision.rwth-aachen.de

Advanced Machine Learning, Winter'15

RWTH AACHEN UNIVERSITY

This Lecture: Advanced Machine Learning

- Regression Approaches
 - Linear Regression
 - Regularization (Ridge, Lasso)
 - Gaussian Processes
- Learning with Latent Variables
 - EM and Generalizations
 - Approximate Inference
- Deep Learning
 - Neural Networks
 - CNNs, RNNs, RBMs, etc.

B. Leibe

RWTH AACHEN UNIVERSITY

Topics of This Lecture

- Recap: Important Concepts from ML Lecture
 - Probability Theory
 - Bayes Decision Theory
 - Maximum Likelihood Estimation
 - Bayesian Estimation
- A Probabilistic View on Regression
 - Least-Squares Estimation as Maximum Likelihood
 - Predictive Distribution
 - Maximum-A-Posteriori (MAP) Estimation
 - Bayesian Curve Fitting
- Discussion

B. Leibe

Advanced Machine Learning, Winter'15

RWTH AACHEN UNIVERSITY

Recap: The Rules of Probability

- Basic rules

Sum Rule $p(X) = \sum_Y p(X, Y)$

Product Rule $p(X, Y) = p(Y|X)p(X)$
- From those, we can derive

Bayes' Theorem $p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$

where $p(X) = \sum_Y p(X|Y)p(Y)$

B. Leibe

Advanced Machine Learning, Winter'15

RWTH AACHEN UNIVERSITY

Recap: Bayes Decision Theory

- Concept 1: Priors (a priori probabilities) $p(C_k)$
 - What we can tell about the probability *before seeing the data*.
 - Example:

a a b a b a a b a
 b a a a b a a b a
 a b a a a b b a
 b a b a a b a a

$P(a)=0.75$
 $P(b)=0.25$

?

$C_1 = a \quad p(C_1) = 0.75$
 $C_2 = b \quad p(C_2) = 0.25$

- In general: $\sum_k p(C_k) = 1$

B. Leibe

Advanced Machine Learning, Winter'15

RWTH AACHEN UNIVERSITY

Recap: Bayes Decision Theory

- Concept 2: Conditional probabilities $p(x|C_k)$
 - Let x be a feature vector.
 - x measures/describes certain properties of the input.
 - E.g. number of black pixels, aspect ratio, ...
 - $p(x|C_k)$ describes its **likelihood** for class C_k .

$p(x|a)$

$p(x|b)$

B. Leibe

Advanced Machine Learning, Winter'15

RWTH AACHEN UNIVERSITY

Recap: Bayes Decision Theory

- **Concept 3: Posterior probabilities** $p(C_k | x)$
 - We are typically interested in the *a posteriori* probability, i.e. the probability of class C_k given the measurement vector x .
- **Bayes' Theorem:**

$$p(C_k | x) = \frac{p(x | C_k) p(C_k)}{p(x)} = \frac{p(x | C_k) p(C_k)}{\sum_i p(x | C_i) p(C_i)}$$
- **Interpretation**

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Normalization Factor}}$$

Slide credit: Bernt Schiele B. Leibe 7

RWTH AACHEN UNIVERSITY

Recap: Bayes Decision Theory

Likelihood

Likelihood \times Prior

Decision boundary

Posterior = $\frac{\text{Likelihood} \times \text{Prior}}{\text{Normalization Factor}}$

Slide credit: Bernt Schiele B. Leibe 8

RWTH AACHEN UNIVERSITY

Recap: Gaussian (or Normal) Distribution

- **One-dimensional case**
 - Mean μ
 - Variance σ^2
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$
- **Multi-dimensional case**
 - Mean μ
 - Covariance Σ
$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}$$

Slide credit: B. Leibe Image source: C. M. Bishop, 2006 9

RWTH AACHEN UNIVERSITY

Side Note

- **Notation**
 - In many situations, it will be preferable to work with the inverse of the **covariance matrix** Σ :
$$\Lambda = \Sigma^{-1}$$
 - We call Λ the **precision matrix**.
 - We can therefore also write the Gaussian as
$$\mathcal{N}(x|\mu, \Lambda^{-1}) = \frac{1}{\sqrt{2\pi}\lambda^{-1/2}} \exp\left\{-\frac{\lambda}{2}(x-\mu)^2\right\}$$

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

Slide credit: B. Leibe 10

RWTH AACHEN UNIVERSITY

Recap: Parametric Methods

- **Given**
 - Data $X = \{x_1, x_2, \dots, x_N\}$
 - Parametric form of the distribution with parameters θ
 - E.g. for Gaussian distrib.: $\theta = (\mu, \sigma)$
- **Learning**
 - Estimation of the parameters θ
- **Likelihood of θ**
 - Probability that the data X have indeed been generated from a probability density with parameters θ

$$L(\theta) = p(X|\theta)$$

Slide adapted from Bernt Schiele B. Leibe 11

RWTH AACHEN UNIVERSITY

Recap: Maximum Likelihood Approach

- **Computation of the likelihood**
 - Single data point: $p(x_n|\theta) = \mathcal{N}(x_n|\mu, \sigma^2)$
 - Assumption: all data points $X = \{x_1, \dots, x_n\}$ are independent
$$L(\theta) = p(X|\theta) = \prod_{n=1}^N p(x_n|\theta)$$
 - Log-likelihood
$$E(\theta) = -\ln L(\theta) = -\sum_{n=1}^N \ln p(x_n|\theta)$$
- **Estimation of the parameters θ (Learning)**
 - Maximize the likelihood (=minimize the negative log-likelihood)
 - ➔ Take the derivative and set it to zero.
$$\frac{\partial}{\partial \theta} E(\theta) = -\sum_{n=1}^N \frac{\partial}{\partial \theta} \ln p(x_n|\theta) \stackrel{!}{=} 0$$

Slide credit: Bernt Schiele B. Leibe 12

RWTH AACHEN UNIVERSITY

Recap: Maximum Likelihood Approach

- Applying ML to estimate the parameters of a Gaussian, we obtain

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N x_n \quad \text{"sample mean"}$$
- In a similar fashion, we get

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \hat{\mu})^2 \quad \text{"sample variance"}$$
- $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$ is the **Maximum Likelihood estimate** for the parameters of a Gaussian distribution.
- This is a very important result.
- Unfortunately, it is biased...

13

RWTH AACHEN UNIVERSITY

Recap: Maximum Likelihood - Limitations

- Maximum Likelihood has several significant limitations
 - It systematically underestimates the variance of the distribution!
 - E.g. consider the case

$$N = 1, X = \{x_1\}$$

⇒ Maximum-likelihood estimate:

- We say ML *overfits to the observed data*.
- We will still often use ML, but it is important to know about this effect.

14

RWTH AACHEN UNIVERSITY

Recap: Deeper Reason

- Maximum Likelihood is a **Frequentist** concept
 - In the **Frequentist view**, probabilities are the frequencies of random, repeatable events.
 - These frequencies are fixed, but can be estimated more precisely when more data is available.
- This is in contrast to the **Bayesian** interpretation
 - In the **Bayesian view**, probabilities quantify the uncertainty about certain states or events.
 - This uncertainty can be revised in the light of new evidence.
- Bayesians and Frequentists do not like each other too well...

15

RWTH AACHEN UNIVERSITY

Recap: Bayesian Approach to Learning

- Conceptual shift
 - Maximum Likelihood views the true parameter vector θ to be unknown, but fixed.
 - In Bayesian learning, we consider θ to be a random variable.
- This allows us to use knowledge about the parameters θ
 - i.e. to use a prior for θ
 - Training data then converts this prior distribution on θ into a posterior probability density.

- The prior thus encodes knowledge we have about the type of distribution we expect to see for θ .

16

RWTH AACHEN UNIVERSITY

Recap: Bayesian Learning Approach

- Bayesian view:**
 - Consider the parameter vector θ as a random variable.
 - When estimating the parameters, what we compute is

$$p(x|X) = \int p(x, \theta|X) d\theta$$

Assumption: given θ , this doesn't depend on X anymore

$$p(x, \theta|X) = p(x|\theta) p(\theta|X)$$

$$p(x|X) = \int p(x|\theta) p(\theta|X) d\theta$$

This is entirely determined by the parameter θ (i.e. by the parametric form of the pdf).

17

RWTH AACHEN UNIVERSITY

Recap: Bayesian Learning Approach

$$p(x|X) = \int p(x|\theta) p(\theta|X) d\theta$$

$$p(\theta|X) = \frac{p(x|\theta) p(\theta)}{p(X)} = \frac{p(\theta)}{p(X)} L(\theta)$$

$$p(X) = \int p(x|\theta) p(\theta) d\theta = \int L(\theta) p(\theta) d\theta$$

- Inserting this above, we obtain

$$p(x|X) = \int \frac{p(x|\theta) L(\theta) p(\theta)}{\int L(\theta) p(\theta) d\theta} d\theta$$

18

RWTH AACHEN UNIVERSITY

Recap: Bayesian Learning Approach

- Discussion
 - Likelihood of the parametric form θ given the data set X .
 - Prior for the parameters θ
 - Estimate for x based on parametric form θ

$$p(x|X) = \int \frac{\underbrace{\overbrace{p(x|\theta)L(\theta)}^{\text{Likelihood of the parametric form } \theta}}{\underbrace{\int L(\theta)p(\theta)d\theta}_{\text{Normalization: integrate over all possible values of } \theta}} p(\theta) d\theta$$

The more uncertain we are about θ , the more we average over all possible parameter values.

19

RWTH AACHEN UNIVERSITY

Topics of This Lecture

- Recap: Important Concepts from ML Lecture
 - Probability Theory
 - Bayes Decision Theory
 - Maximum Likelihood Estimation
 - Bayesian Estimation
- A Probabilistic View on Regression
 - Least-Squares Estimation as Maximum Likelihood
 - Predictive Distribution
 - Maximum-A-Posteriori (MAP) Estimation
 - Bayesian Curve Fitting
- Discussion

20

RWTH AACHEN UNIVERSITY

Curve Fitting Revisited

- In the last lecture, we've looked at curve fitting in terms of error minimization...
- Now: View the problem from a probabilistic perspective
 - Goal is to make predictions for target variable t given new value for input variable x .
 - Basis: training set $\mathbf{x} = (x_1, \dots, x_N)^T$ with target values $\mathbf{t} = (t_1, \dots, t_N)^T$.
 - We express our uncertainty over the value of the target variable using a probability distribution

$$p(t|x, \mathbf{w}, \beta)$$

21

RWTH AACHEN UNIVERSITY

Probabilistic Regression

- First assumption:
 - Our target function values t are generated by adding noise to the ideal function estimate:

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

Target function value (points to t) ← *Regression function* (points to $y(\mathbf{x}, \mathbf{w})$) ← *Input value* (points to \mathbf{x}) ← *Weights or parameters* (points to \mathbf{w}) ← *Noise* (points to ϵ)
- Second assumption:
 - The noise is Gaussian distributed.

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

Mean (points to $y(\mathbf{x}, \mathbf{w})$) ← *Variance (β precision)* (points to β^{-1})

22

RWTH AACHEN UNIVERSITY

Visualization: Gaussian Noise

23

RWTH AACHEN UNIVERSITY

Probabilistic Regression

- Given
 - Training data points: $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$
 - Associated function values: $\mathbf{t} = [t_1, \dots, t_n]^T$
- Conditional likelihood (assuming i.i.d. data)

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n|y(\mathbf{x}_n, \mathbf{w}), \beta^{-1}) = \prod_{n=1}^N \mathcal{N}(t_n|\mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1})$$

Generalized linear regression function (points to $\mathbf{w}^T \phi(\mathbf{x}_n)$)

\Rightarrow Maximize w.r.t. \mathbf{w}, β

24

RWTH AACHEN UNIVERSITY

Maximum Likelihood Regression

- Simplify the log-likelihood

$$\log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \sum_{n=1}^N \log \mathcal{N}(t_n | y(\mathbf{x}_n, \mathbf{w}), \beta^{-1})$$

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

$$= \sum_{n=1}^N \left[\log\left(\frac{\sqrt{\beta}}{\sqrt{2\pi}}\right) - \frac{\beta}{2} \{y(\mathbf{x}_n, \mathbf{w}) - t_n\}^2 \right]$$

$$= -\frac{\beta}{2} \sum_{n=1}^N \underbrace{\{t_n - y(\mathbf{x}_n, \mathbf{w})\}^2}_{\text{Sum-of-squares error}} + \underbrace{\frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi)}_{\text{Constants}}$$

Slide adapted from Bernt Schiele. B. Leibe 25

RWTH AACHEN UNIVERSITY

Maximum Likelihood Regression

$$\log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^N \{t_n - y(\mathbf{x}_n, \mathbf{w})\}^2 + \frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi)$$

$$= -\frac{\beta}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 + \frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi)$$

- Gradient w.r.t. \mathbf{w} :

$$\nabla_{\mathbf{w}} \log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = -\beta \sum_{n=1}^N (t_n - \mathbf{w}^T \phi(\mathbf{x}_n)) \phi(\mathbf{x}_n)$$

B. Leibe 26

RWTH AACHEN UNIVERSITY

Maximum Likelihood Regression

$$\nabla_{\mathbf{w}} \log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = -\beta \sum_{n=1}^N (t_n - \mathbf{w}^T \phi(\mathbf{x}_n)) \phi(\mathbf{x}_n)$$

- Setting the gradient to zero:

$$0 = -\beta \sum_{n=1}^N (t_n - \mathbf{w}^T \phi(\mathbf{x}_n)) \phi(\mathbf{x}_n)$$

$$\Leftrightarrow \sum_{n=1}^N t_n \phi(\mathbf{x}_n) = \left[\sum_{n=1}^N \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T \right] \mathbf{w}$$

$$\Leftrightarrow \Phi \mathbf{t} = \Phi \Phi^T \mathbf{w} \quad \Phi = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_N)]$$

$$\Leftrightarrow \mathbf{w}_{\text{ML}} = (\Phi \Phi^T)^{-1} \Phi \mathbf{t} \quad \leftarrow \text{Same as in least-squares regression!}$$

\Rightarrow Least-squares regression is equivalent to Maximum Likelihood under the assumption of Gaussian noise.

Slide adapted from Bernt Schiele. B. Leibe 28

RWTH AACHEN UNIVERSITY

Role of the Precision Parameter

- Also use ML to determine the precision parameter β :

$$\log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 + \frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi)$$

- Gradient w.r.t. β :

$$\nabla_{\beta} \log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = -\frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 + \frac{N}{2} \frac{1}{\beta}$$

$$\frac{1}{\beta_{\text{ML}}} = \frac{1}{N} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2$$

\Rightarrow The inverse of the noise precision is given by the residual variance of the target values around the regression function.

B. Leibe 29

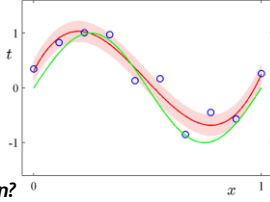
RWTH AACHEN UNIVERSITY

Predictive Distribution

- Having determined the parameters \mathbf{w} and β , we can now make predictions for new values of \mathbf{x} .

$$p(t|\mathbf{X}, \mathbf{w}_{\text{ML}}, \beta_{\text{ML}}) = \mathcal{N}(t | y(\mathbf{x}, \mathbf{w}_{\text{ML}}), \beta_{\text{ML}}^{-1})$$

- This means
 - Rather than giving a point estimate, we can now also give an estimate of the estimation uncertainty.



- What else can we do in the Bayesian view of regression?

B. Leibe. Image source: C. M. Bishop, 2006. 30

RWTH AACHEN UNIVERSITY

MAP: A Step Towards Bayesian Estimation...

- Introduce a prior distribution over the coefficients \mathbf{w} .
 - For simplicity, assume a zero-mean Gaussian distribution
$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2} \mathbf{w}^T \mathbf{w}\right\}$$
 - New hyperparameter α controls the distribution of model parameters.
- Express the posterior distribution over \mathbf{w} .
 - Using Bayes' theorem:
$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \beta, \alpha) \propto p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha)$$
 - We can now determine \mathbf{w} by maximizing the posterior.
 - This technique is called **maximum-a-posteriori (MAP)**.

B. Leibe 31

RWTH AACHEN UNIVERSITY

MAP Solution

- Minimize the negative logarithm
 - $-\log p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \beta, \alpha) \propto -\log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) - \log p(\mathbf{w}|\alpha)$
 - $-\log p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \frac{\beta}{2} \sum_{n=1}^N \{y(\mathbf{x}_n, \mathbf{w}) - t_n\}^2 + \text{const}$
 - $-\log p(\mathbf{w}|\alpha) = \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} + \text{const}$
- The MAP solution is therefore the solution of
 - $\frac{\beta}{2} \sum_{n=1}^N \{y(\mathbf{x}_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w}$
 - \Rightarrow Maximizing the posterior distribution is equivalent to minimizing the regularized sum-of-squares error (with $\lambda = \frac{\alpha}{\beta}$).

B. Leibe

RWTH AACHEN UNIVERSITY

Results of Probabilistic View on Regression

- Better understanding what linear regression means
 - Least-squares regression is equivalent to ML estimation under the assumption of Gaussian noise.
 - \Rightarrow We can use the predictive distribution to give an uncertainty estimate on the prediction.
 - \Rightarrow But: known problem with ML that it tends towards overfitting.
 - L2-regularized regression (Ridge regression) is equivalent to MAP estimation with a Gaussian prior on the parameters \mathbf{w} .
 - \Rightarrow The prior controls the parameter values to reduce overfitting.
 - \Rightarrow This gives us a tool to explore more general priors.
- But still no full Bayesian Estimation yet
 - Should integrate over all values of \mathbf{w} instead of just making a point estimate.

B. Leibe

RWTH AACHEN UNIVERSITY

Bayesian Curve Fitting

- Given
 - Training data points: $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$
 - Associated function values: $\mathbf{t} = [t_1, \dots, t_n]^T$
 - Our goal is to predict the value of t for a new point x .
- Evaluate the predictive distribution
 - $$p(t|x, \mathbf{X}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{X}, \mathbf{t}) d\mathbf{w}$$
 - What we just computed for MAP
 - Noise distribution - again assume a Gaussian here
 - $p(t|x, \mathbf{w}) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1})$
 - Assume that parameters α and β are fixed and known for now.

B. Leibe

RWTH AACHEN UNIVERSITY

Bayesian Curve Fitting

- Under those assumptions, the posterior distribution is a Gaussian and can be evaluated analytically:
 - $$p(t|x, \mathbf{X}, \mathbf{t}) = \mathcal{N}(t|m(x), s^2(x))$$
 - where the mean and variance are given by
 - $$m(x) = \beta \phi(x)^T \mathbf{S} \sum_{n=1}^N \phi(\mathbf{x}_n) t_n$$
 - $$s(x)^2 = \beta^{-1} + \phi(x)^T \mathbf{S} \phi(x)$$
 - and \mathbf{S} is the regularized covariance matrix
 - $$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^N \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T$$

B. Leibe

RWTH AACHEN UNIVERSITY

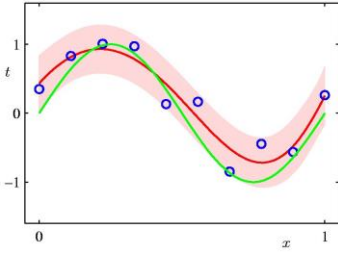
Analyzing the result

- Analyzing the variance of the predictive distribution
 - $$s(x)^2 = \beta^{-1} + \phi(x)^T \mathbf{S} \phi(x)$$
 - Uncertainty in the predicted value due to noise on the target variables (expressed already in ML)
 - Uncertainty in the parameters \mathbf{w} (consequence of Bayesian treatment)

B. Leibe

RWTH AACHEN UNIVERSITY

Bayesian Predictive Distribution



- Important difference to previous example
 - Uncertainty may vary with test point x !

B. Leibe

Image source: C. M. Bishop, 2006

RWTH AACHEN
UNIVERSITY

Topics of This Lecture

- Recap: Important Concepts from ML Lecture
 - Probability Theory
 - Bayes Decision Theory
 - Maximum Likelihood Estimation
 - Bayesian Estimation
- A Probabilistic View on Regression
 - Least-Squares Estimation as Maximum Likelihood
 - Predictive Distribution
 - Maximum-A-Posteriori (MAP) Estimation
 - Bayesian Curve Fitting
- Discussion

38

RWTH AACHEN
UNIVERSITY

Discussion

- We now have a better understanding of regression
 - Least-squares regression: Assumption of Gaussian noise
 - ⇒ We can now also plug in different noise models and explore how they affect the error function.
 - L2 regularization as a Gaussian prior on parameters w .
 - ⇒ We can now also use different regularizers and explore what they mean.
 - ⇒ Next lecture...
 - General formulation with basis functions $\phi(x)$.
 - ⇒ We can now also use different basis functions.

39

RWTH AACHEN
UNIVERSITY

Discussion

- General regression formulation
 - In principle, we can perform regression in arbitrary spaces and with many different types of basis functions
 - However, there is a caveat... Can you see what it is?
- Example: Polynomial curve fitting, $M = 3$

$$y(x, \mathbf{w}) = w_0 + \sum_{i=1}^D w_i x_i + \sum_{i=1}^D \sum_{j=1}^D w_{ij} x_i x_j + \sum_{i=1}^D \sum_{j=1}^D \sum_{k=1}^D w_{ijk} x_i x_j x_k$$
 - ⇒ Number of coefficients grows with D^M
 - ⇒ The approach becomes quickly unpractical for high dimensions.
 - This is known as the **curse of dimensionality**.
 - We will encounter some ways to deal with this later.

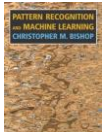
40

RWTH AACHEN
UNIVERSITY

References and Further Reading

- More information on linear regression can be found in Chapters 1.2, 5-1.2.6 and 3.1-3.1.4 of

Christopher M. Bishop
Pattern Recognition and Machine Learning
Springer, 2006



41