Linear Models for Regression

Henrik I Christensen

Robotics & Intelligent Machines @ GT
Georgia Institute of Technology,
Atlanta, GA 30332-0280
hic@cc.gatech.edu

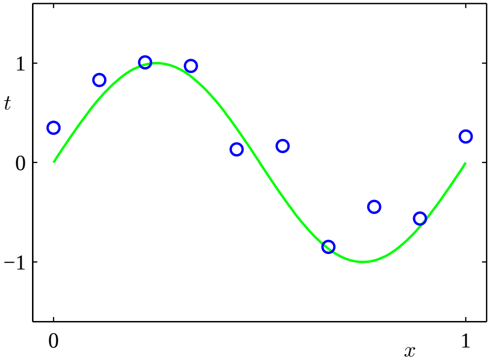
Outline

Introduction

- The objective of regression is to enable prediction of a value t based on modelling over a dataset X.
- Consider a set of D observations over a space
- How can we generate estimates for the future?
  - Battery time?
  - Time to completion?
  - Position of doors?

Introduction (2)

- Example from Chapter 1



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_m x^m = \sum_{i=0}^{m} w_i x^i$$

Introduction (3)

- In general the functions could be beyond simple polynomials
- The "components" are termed *basis functions*, i.e.

$$y(x, \mathbf{w}) = \sum_{i=0}^{m} w_i \phi_i(x) = \vec{w}^T \phi(\vec{x})$$

Outline

Loss Function

- For optimization we need a penalty / loss function

$$L(t, y(x))$$

- Expected loss is then

$$E[L] = \int\int L(t, y(x))p(x, t)dxdt$$

- For the squared loss function we have

$$E[L] = \int\int \{y(x) - t\}^2 p(x, t)dxdt$$

- Goal: choose y(x) to minimize expected loss (E[L])

Loss Function

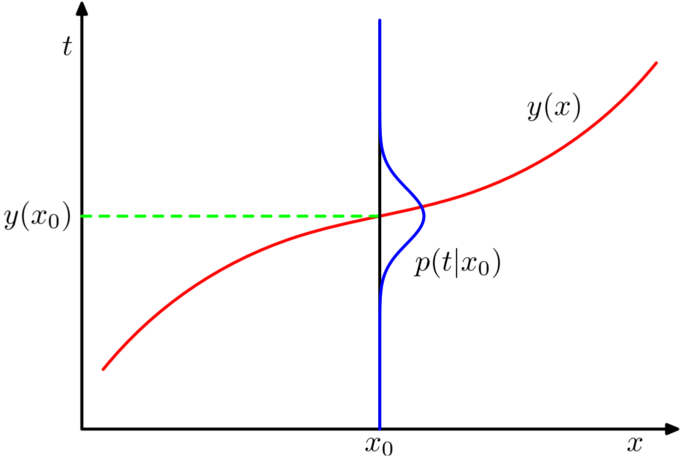
- Derivation of the extrema

$$\frac{\delta E[L]}{\delta y(x)} = 2 \int \{y(x) - t\} p(x,t) dt = 0$$

- Implies that

$$y(x) = \frac{\int tp(x,t)dt}{p(x)} = \int tp(t|x)dt = E[t|x]$$

Loss Function - Interpretation



Alternative

- Consider a small rewrite

$$\{y(x) - t\}^2 = \{y(x) - E[t|x] + E[t|x] - t\}^2$$

- The expected loss is then

$$E[L] = \int \{y(x) - E[t|x]\}^2 p(x) dx + \int \{E[t|x] - t\}^2 p(x) dx$$

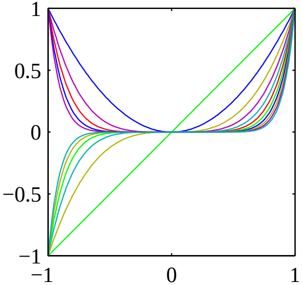
Outline

Polynomial Basis Functions

Basic Definition:

$$\phi_i(x) = x^i$$

Global functions
Small change in x affects all of them



Gaussian Basis Functions

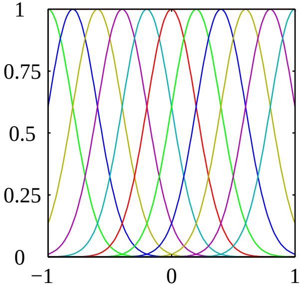
Basic Definition:

$$\phi_i(x) = e^{-\frac{(x-\mu_i)^2}{2s^2}}$$

A way to Gaussian mixtures, local impact

Not required to have probabilistic interpretation.

μ control position and s control scale



Sigmoid Basis Functions

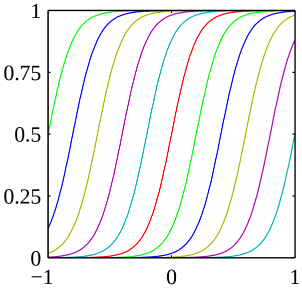
Basic Definition:

$$\phi_i(x) = \sigma\left(\frac{x - \mu_i}{s}\right)$$

where

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

μ controls location and s controls slope



Maximum Likelihood & Least Squares

- Assume observation from a deterministic function contaminated by Gaussian Noise

$$t = y(x, w) + \epsilon \quad p(\epsilon|\beta) = N(\epsilon|0, \beta^{-1})$$

the problem at hand is then

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

- From a series of observations we have the likelihood

$$p(\mathbf{t}|\mathbf{X}|w, \beta) = \prod_{i=1}^{N} N(t_i|w^T\phi(x_i), \beta^{-1})$$

Maximum Likelihood & Least Squares (2)

- This results in

$$\ln p(\mathbf{t}|\mathbf{w}, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w})$$

- where

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} \{t_i - \mathbf{w}^T \phi(x_i)\}^2$$

is the sum of squared errors

Maximum Likelihood & Least Squares (3)

- Computing the extrema yields:

$$\mathbf{w}_{ML} = \left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{t}$$

- where

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{M-1}(x_1) \\ \phi_0(x_1) & \phi_1(x_2) & \cdots & \phi_{M-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & \cdots & \phi_{M-1}(x_N) \end{pmatrix}$$

Line Estimation

- Least square minimization:
  - Line equation: $y = ax + b$
  - Error in fit: $\sum_i (y_i - ax_i - b)^2$
  - Solution:

$$\begin{pmatrix} \bar{y^2} \\ \bar{y} \end{pmatrix} = \begin{pmatrix} \bar{x^2} & \bar{x} \\ \bar{x} & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

  - Minimizes vertical errors. Non-robust!

LSQ on Lasers

- Line model: $r_i \cos(\phi_i - \theta) = \rho$
- Error model: $d_i = r_i \cos(\phi_i - \theta) - \rho$
- Optimize: $\mathrm{argmin}_{(\rho,\theta)} \sum_i (r_i \cos(\phi_i - \theta) - \rho)^2$
- Error model derived in Deriche *et al.* (1992)
- Well suited for "clean-up" of Hough lines

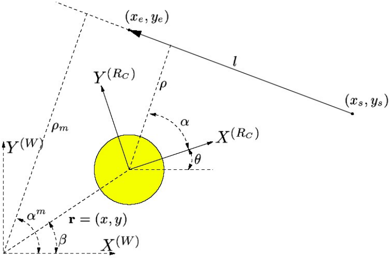
Total Least Squares

- Line equation: $ax + by + c = 0$
- Error in fit: $\sum_i (ax_i + by_i + c)^2$ where $a^2 + b^2 = 1$.
- Solution:

$$\begin{pmatrix} \overline{x^2} - \bar{x}\bar{x} & \overline{xy} - \bar{x}\bar{y} \\ \overline{xy} - \bar{x}\bar{y} & \overline{y^2} - \bar{y}\bar{y} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \mu \begin{pmatrix} a \\ b \end{pmatrix}$$

  where μ is a scale factor.
- $c = -a\bar{x} - b\bar{y}$

Line Representations



- The line representation is crucial
- Often a redundant model is adopted
- Line parameters vs end-points
- Important for fusion of segments.
- End-points are less stable

Sequential Adaptation

- In some cases one at a time estimation is more suitable
- Also known as gradient descent

$$\begin{aligned} \mathbf{w}^{(\tau+1)} &= \mathbf{w}^{(\tau)} - \eta \nabla E_n \\ &= \mathbf{w}^{(\tau)} - \eta(t_n - \mathbf{w}^{(\tau)T}\phi(x_n))\phi(x_n) \end{aligned}$$

- Knows as least-mean square (LMS). An issue is how to choose η?

Regularized Least Squares

- As seen in lecture 2 sometime control of parameters might be useful.
- Consider the error function:

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

- which generates

$$\frac{1}{2}\sum_{i=1}^{N}\{t_i - w^t\phi(x_i)\}^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

- which is minimized by

$$w = \left(\lambda I + \Phi^T\Phi\right)^{-1}\Phi^T\mathbf{t}$$

Outline

Bayesian Linear Regression

- Define a conjugate prior over w

$$p(w) = N(w|m_0, S_0)$$

- given the likelihood function and regular from Bayesian analysis we can derive

$$p(w|t) = N(w|m_N, S_N)$$

- where

$$\begin{aligned} m_N &= S_N\left(S_0^{-1}m_0 + \beta\Phi^T t\right) \\ S_N^{-1} &= S_0^{-1} + \beta\Phi^T\Phi \end{aligned}$$

Bayesian Linear Regression (2)

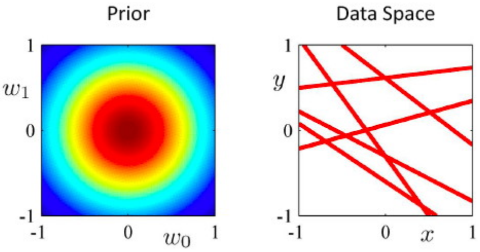
- A common choice is

$$p(w) = N(w|0, \alpha^{-1} I)$$

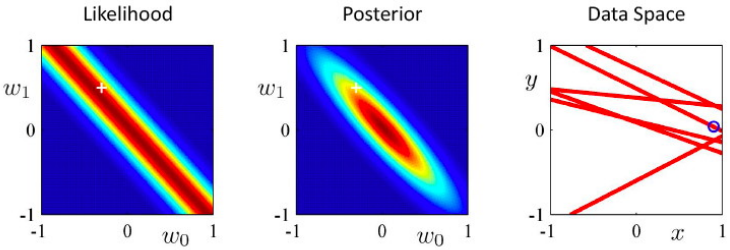
- So that

$$\begin{aligned} m_N &= \beta S_N \Phi^T t \\ S_N^{-1} &= \alpha I + \beta \Phi^T \Phi \end{aligned}$$

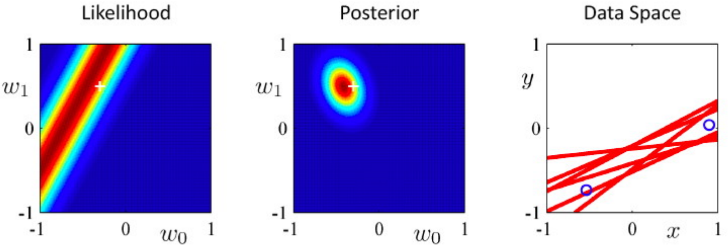
Example - No Data



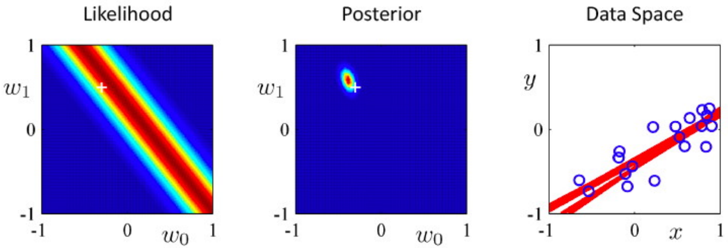
Example - 1 Data Point



Example - 2 Data Points



Example - 20 Data Points



Outline

Bayesian Model Comparison

- How does one select an appropriate model?
- Assume for a minute we want to compare a set of models M_i, $i \in 1, \ldots L$ for a dataset D
- We could compute

$$p(M_i|D) \propto p(D|M_i)p(M_i)$$

- Bayes Factor: Ratio of evidence for two models

$$\frac{p(D|M_i)}{p(D|M_j)}$$

The mixture distribution approach

- We could use all the models:

$$p(t|x, D) = \sum_{i=1}^{L} p(t|x, M_i, D)p(M_i|D)$$

- Or simply go with the most probably/best model.

Model Evidence

- We can compute model evidence

$$p(D|M_i) = \int p(D|w, M_i)p(w|M_i)dw$$

- Allow computation of model fit based on parameter range

Evaluation of Parameters

- Evaluation of posterior over parameters

$$p(w|D, M_i) = \frac{P(D|w, M_i)p(w|M_i)}{P(D|M_i)}$$

- There is a need to understand how good is a model?

Model Comparison

- Consider evaluation of a model w. parameters w

$$p(D) = \int p(D|w)p(w)dw \approx p(D|w_{map})\frac{\sigma_{posterior}}{\sigma_{prior}}$$

- Then

$$\ln p(D) \approx \ln p(D|w_{map}) + \ln\left(\frac{\sigma_{posterior}}{\sigma_{prior}}\right)$$

Model Comparison as Kullback-Leibler

- From earlier we have comparison of distributions

$$KL = \int p(D|M_1) \ln \frac{p(D|M_1)}{p(D|M_2)} dD$$

- Enables comparison of two different models

Outline

Summary

- Brief intro to linear methods for estimation of models
- Prediction of values and models
  - Needed for adaptive selection of models (black-box/grey-box)
  - Evaluation of sensor models, ...
- Consideration of batch and recursive estimation methods
- Significant discussion of methods for evaluation of models and parameters.
- This far purely a discussion of linear models

Deriche, R., Vaillant, R., & Faugeras, O. 1992. *From Noisy Edges Points to 3D Reconstruction of a Scene : A Robust Approach and Its Uncertainty Analysis*. Vol. 2. World Scientific. Series in Machine Perception and Artificial Intelligence. Pages 71–79.