Regression

Reading: Chapter 8.1-8.4 Prince Book

Outline

- Linear regression
- Bayesian regression
- Non-linear regression
- Kernel “trick”
- Gaussian process regression

Linear Regression

We have one equation for each x,w training pair:

\[ Pr(w_i | x_i, \theta) = \text{Norm}_{\phi} \left( \phi^T x_i, \sigma^2 \right) \]

Joint Likelihood over whole training dataset

\[ Pr(w) = \text{Norm}_{\phi} [X^T \phi, \sigma^2 I] \]

where

\[ X = [x_1, x_2 \ldots x_n] \quad w = [w_1, w_2, \ldots, w_n]^T \]

Bayesian Regression

Likelihood

\[ Pr(w | X) = \text{Norm}_{\phi} [X^T \phi, \sigma^2 I] \]

Prior

\[ Pr(\phi) = \text{Norm}_{\phi} (0, \sigma^2 I) \]

Bayes rule

\[ Pr(\phi | X, w) = \frac{Pr(w | X, \phi) Pr(\phi)}{Pr(w | X)} \]

Bayesian Regression

\[ Pr(\phi | X, w) = \text{Norm}_{\phi} \left[ \frac{1}{\sigma^2} A^{-1} X w, A^{-1} \right] \]

where

\[ A = \frac{1}{\sigma^2} XX^T + \frac{1}{\sigma^2} I \]
Bayesian Regression

\[ P(w|x^*, \phi) \propto \int P(w^*|x^*, \phi) P(\phi|X, w) d\phi \]

\[ = \int \text{Norm}_{\phi} \left[ \phi^T x^* \sigma^2 \text{Norm}_{w} \left[ \frac{1}{2} A^{-1} X w - A^{-1} \right] d\phi \right] \]

\[ = \text{Norm}_{w} \left[ \frac{1}{2} x^T A^{-1} X x - A^{-1} x + \sigma^2 \right] \]

Non-Linear Regression

**GOAL:**
Keep the math of linear regression, but extend to more general functions

**KEY IDEA:**
You can make a non-linear function from a linear weighted sum of non-linear basis functions

Non-linear regression

Linear regression:

\[ P(w|x_i, \theta) = \text{Norm}_{w_i} \left[ \phi^T x_i \sigma^2 \right] \]

Non-Linear regression:

\[ P(w|x_i, \theta) = \text{Norm}_{w_i} \left[ \phi^T z_i \sigma^2 \right] \]

where \( z_i = f[x_i] \)

In other words, create \( z \) by evaluating \( x \) against basis functions, then linearly regress against \( z \).

Example: polynomial regression

\[ P(w|x_i) = \text{Norm}_{w_i} \left[ \phi^T z_i \sigma^2 \right] \]

A special case of

\[ P(w|x_i) = \text{Norm}_{w_i} \left[ \phi^T z_i \sigma^2 \right] \]

Where

\[ z_i = \begin{bmatrix} 1 \\ x_i^2 \\ \ldots \end{bmatrix} \]

Radial basis functions

Arc Tan Functions

**note:** sigmoid-like functions

\[ x_i = \begin{bmatrix} \exp \left[ -(x_i - \alpha_1)^2 / \lambda \right] \\ \exp \left[ -(x_i - \alpha_2)^2 / \lambda \right] \\ \exp \left[ -(x_i - \alpha_3)^2 / \lambda \right] \\ \exp \left[ -(x_i - \alpha_4)^2 / \lambda \right] \end{bmatrix} \]

\[ x_i = \begin{bmatrix} \text{arctan}(x_i - \alpha_1) \\ \text{arctan}(x_i - \alpha_2) \\ \text{arctan}(x_i - \alpha_3) \\ \text{arctan}(x_i - \alpha_4) \end{bmatrix} \]
Nonlinear Regression

Maximum likelihood estimation for nonlinear regression is the same as linear regression, but substitute in $Z$ for $X$:

$$\hat{\phi} = (ZZ^T)^{-1}Zw,$$
$$\hat{\sigma}^2 = \frac{\left(w - Z^T\phi\right)(w - Z^T\phi)}{N}.$$
Gaussian Process Regression
• Bayesian nonlinear regression using kernels!

RBF Kernel Fits

Figure 8.9 Gaussian process regression using an RBF kernel a) When the length scale parameter $\lambda$ is large, the function is too smooth. b) For small values of the length parameter the model does not successfully interpolate between the examples. c) The regression using the maximum likelihood length scale parameter is neither too smooth nor disjointed.

\[ k(x_i, x_j) = \exp\left( -\frac{(x_i - x_j)^2}{\lambda} \right) \]