DM825

Introduction to Machine Learning

Lecture 2

Linear Models and Probabilistic Interpretation

Marco Chiarandini

Department of Mathematics & Computer Science
University of Southern Denmark

Outline

Outline

- Linear Regression
- k Nearest Neighbor
- Curse of Dimensionality

Outline

Linear Models

We saw linear combination of input variables. We can generalize to other functions while preserving linearity in $\vec{\theta}$, eg, polynomials.

$\rightsquigarrow$ linear combination of a fixed set of nonlinear functions of input variables known as basis functions.

$(\vec{y}, \vec{x})$ training data $(\hat{y}, h_\theta(\vec{x}))$ prediction on new data

$h(\vec{x}, \vec{\theta}) = \theta_0 + \theta_1 x_1 + \ldots + \theta_p x_p$ linear regression

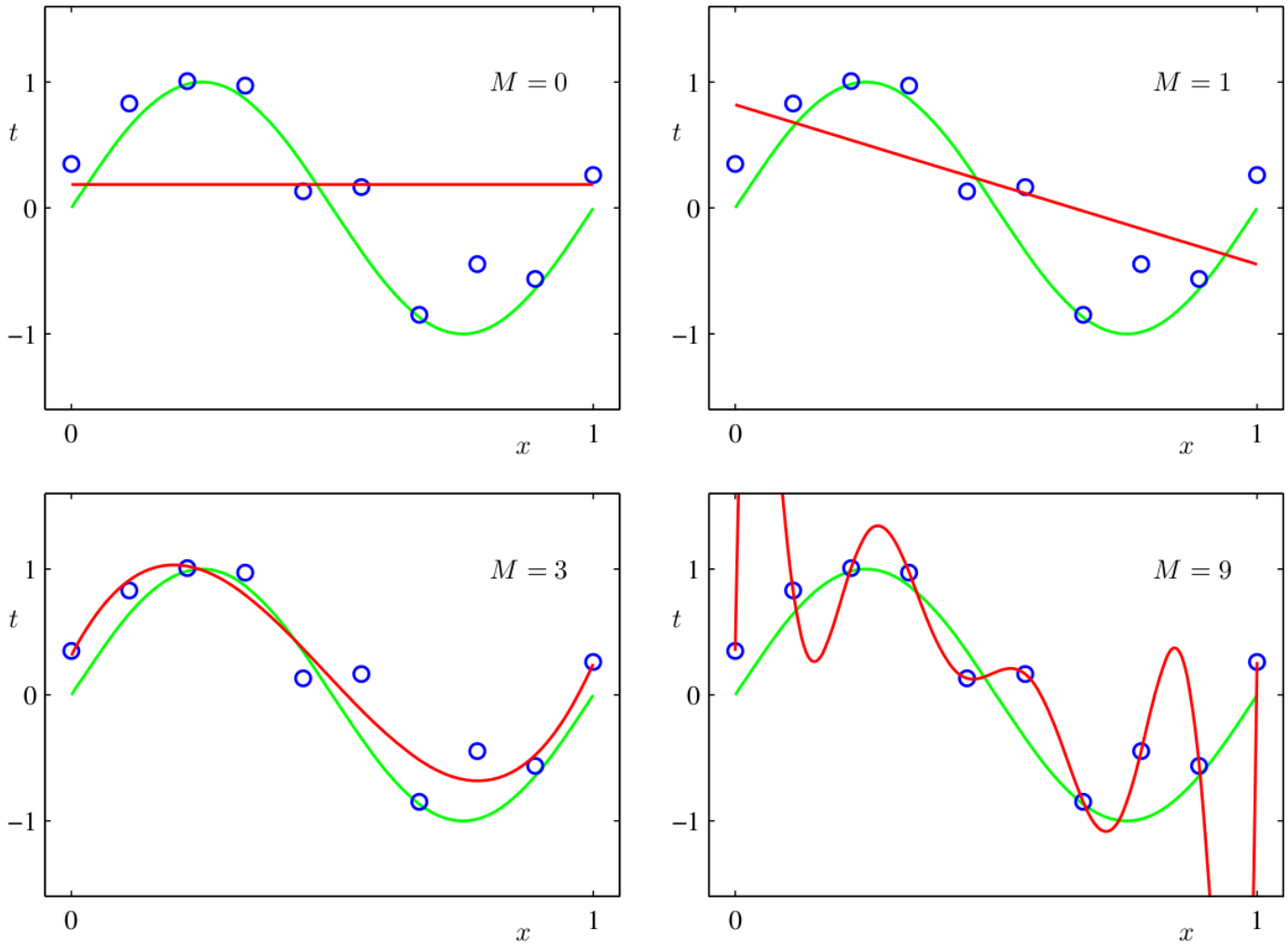
$$h(\vec{x}, \vec{\theta}) = \theta_0 + \sum_{j=1}^{p} \theta_j x_j + \sum_{i=1}^{p} \sum_{j=1}^{p} \theta_{ij} x_i x_j + \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{k=1}^{p} \theta_{ijk} x_i x_j x_k \text{ polynomial}$$

$$h(\vec{x}, \vec{\theta}) = \theta_0 + \sum_{j=1}^{p} \theta_j \phi_j(\vec{x}) = \vec{\theta}^T \vec{\phi}(\vec{x}) \text{ linear models}$$

h is now a nonlinear function of input vector $\vec{x}$ but h is linear in $\vec{\theta}$

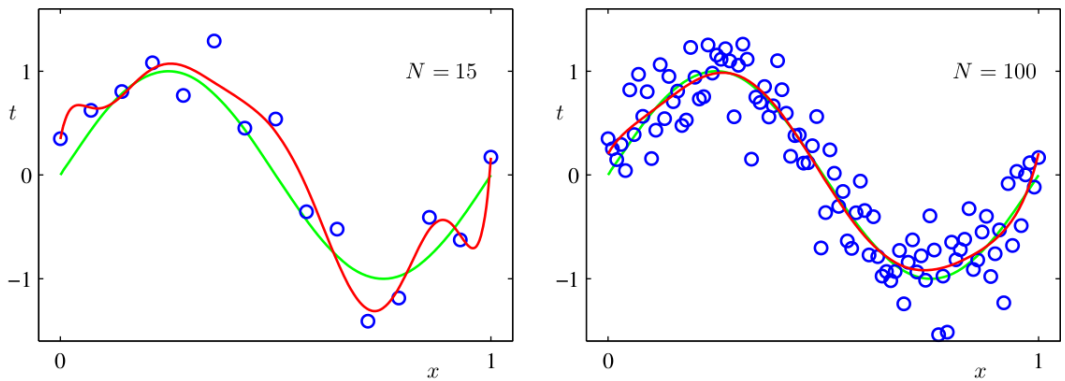
even though, parameters remain easy to estimate, curse of dimensionality

Overfitting



Regularization

- in overfitting parameters θ reach high values
- rule of thumb: 5,10 times more data then parameters



- counteract this by introducing a regularization term in the cost function:

in statistics shrinkage and
ridge regression
in ML weight decay

$$\tilde{L}(\vec{\theta}) = \frac{1}{2} L(\vec{\theta}) + \lambda E_{\vec{\theta}}(\vec{\theta})$$

$$= \frac{1}{2} \sum_{i=1}^{m} \left(y^i - \vec{\theta}^T \vec{\phi}(\vec{x}^i) \right)^2 + \frac{\lambda}{2} \vec{\theta}^T \vec{\theta}$$

It remains a quadratic function that can be solved analytically:

$$\vec{\theta} = (\lambda \mathbf{I} + \boldsymbol{\phi}^T \boldsymbol{\phi})^{-1} \boldsymbol{\phi}^T \vec{y}$$

problem shifted to determine λ

Try in R via `optim`

Locally Weighted Linear Regression

Linear models are global functions so changes in one region affect everywhere

- divide into regions, fit different polynomials in each region $\rightsquigarrow$ spline function

- locally weighted linear regression
  Ordinary lin. regr.: to predict any query point $\vec{x}$ carry out step 2 below:

  1. fit $\vec{\theta} = \text{argmin} \sum_i (y^i - \vec{\theta}^T \vec{x}^i)^2$
  2. output $\vec{\theta}\vec{x}$

  Loc. lin. reg. (nonparametric method): repeat for each $\vec{x}$ to predict:

  1. fit $\vec{\theta} = \text{argmin} \sum_i w_i (y^i - \vec{\theta}^T \vec{x}^i)^2$
  2. output $\vec{\theta}\vec{x}$

  $w^i = \exp\left(-\frac{(\vec{x}^i - \vec{x})^T (\vec{x}^i - \vec{x})}{2\tau^2}\right)$, τ bandwidth

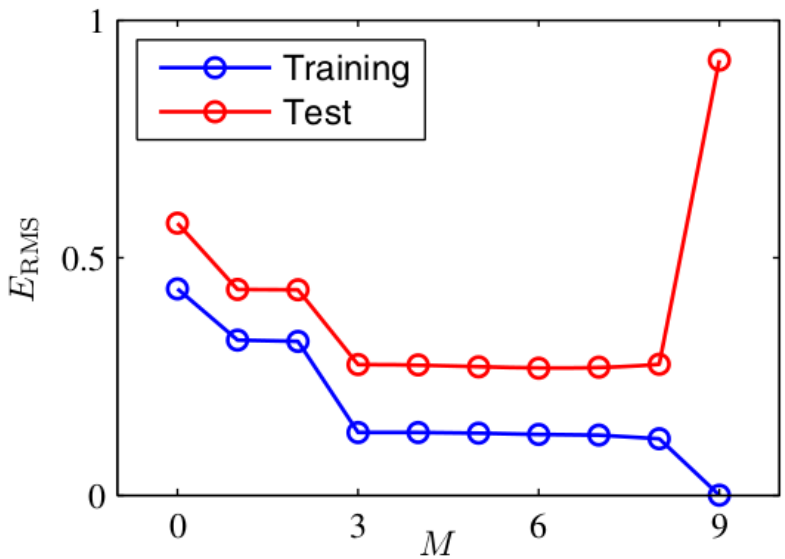
  if $||x^i - x||$ is small $\Longrightarrow w^i$ close to 1; if large $\Longrightarrow w^i$ close to 0;

Model Comparison

$$L(\vec{\theta}) = \frac{1}{2}\left[h(x) - y\right]^2$$ on training data
on test data

$$E[L(\theta)] = \frac{1}{2}\sum_{i=1}^{m}\left[h(x) - y\right]^2$$ average loss

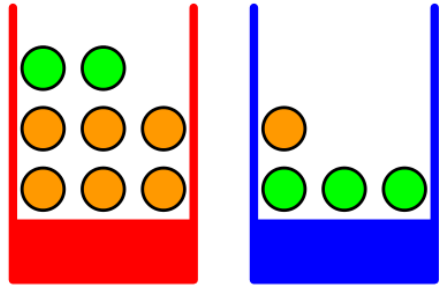
$$E_{RMS} = \sqrt{2E[L(\vec{\theta})]/m}$$ root mean square



Outline

Probability Theory Review

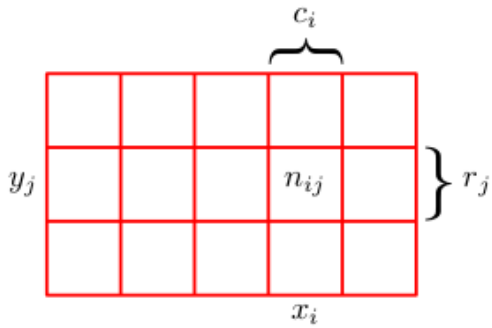
We randomly select one of the boxes and from that box we randomly pick with replacement an item of fruit.



X, Y random variables (the Box and the Fruit)

$x_i, i = 1 \dots M, y_j, j = 1 \dots L$ values

$\Pr(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$ joint probability



$\Pr(X = x_i) = \sum_{j=1}^{L} \Pr(X = x_i, Y = y_j) = \frac{c_i}{N}$ marginal prob. (product rule)

$\Pr(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}$ conditional probability

$\Pr(X = x_i, Y = y_j) = \Pr(Y = y_j \mid X = x_i) \Pr(X = x_i)$ product rule

Note:
$p(X)$ is probability distribution of a random variable
$p(x)$ is the distribution evaluated for that particular value x

$p(X,Y) = p(X)p(Y) \iff p(Y \mid X) = p(Y)$ independency

product rule + $p(X,Y) = p(Y,X) \implies$ Bayes rule:

$$p(Y \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)} = \frac{p(X \mid Y)p(Y)}{\sum_Y p(X \mid Y)p(Y)}$$

Continuous variables

$p(x)$ probability density over x, ie, prob. of falling in $(x, x + \delta x)$

$p(x \in (a,b)) = \int_a^b p(x)dx$ probability density function
(if x is discrete $p(x)$ is probability mass function)

$P(z) = \int_{-\infty}^{z} p(x)dx$ cumulative distribution function

$p(\vec{x}) = p(x_1, \ldots, x_k)$

$p(x) = \int p(x,y)dy$ marginalization (sum rule)

$p(x,y) = p(y|x)p(x)$ product rule

average value of some function $f(x)$ under $p(x)$

$$E[f] = \sum_x p(x)f(x) \qquad E[f] = \int p(x)f(x)dx$$

$E_x[f(x,y)]$ if several variables we specify

$E_x[f|y] = \sum_x p(x|y)f(x)$ conditional expectation with respect to some conditional distribution

$\mathrm{var}[f] = E\left[(f(x) - E[f(x)])^2\right] = E[f(x)^2] - E[f(x)]^2$ variance

Bayesian Probabilities

Classical or frequentist notion frequency of observed values of random variables: relative occurrence of the values
criticism: does not work with unrepeatable events

Bayesian perspective looks at the uncertainty that surrounds the model parameters $\vec{w}$

We capture our assumptions about $\vec{w}$ before observing data in the form of a prior probability distribution $p(\vec{w})$.

$\mathcal{D} = \{y^1, \ldots, y^m\}$ observed data

$p(\mathcal{D}|\vec{w})$ conditional probability or likelihood function
how probable the observed data is for different settings of parameter $\vec{w}$

$p(\vec{w}|\mathcal{D})$ effect of observed data, uncertainty of $\vec{w}$ after observed $\mathcal{D}$, posterior

$$p(\vec{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\vec{w})p(\vec{w})}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\vec{w})p(\vec{w})}{\int p(\mathcal{D}|\vec{w})p(\vec{w})d\vec{w}}$$

posterior $\propto$ likelihood $\times$ prior

Frequentist $\vec{w}$ is a fixed parameter, whose value is determined by
"estimators" and confidence intervals
maximum likelihood method:

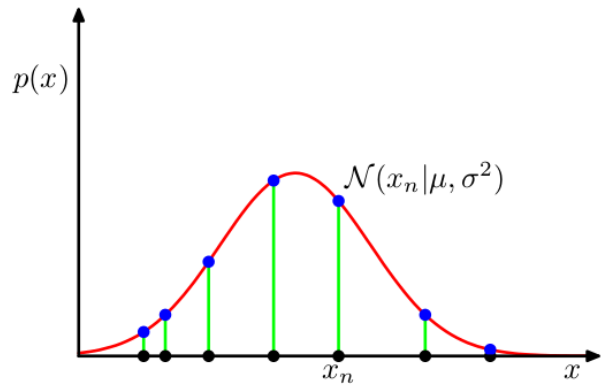
$$\begin{aligned}\vec{w} &= \operatorname{argmax} p(\mathcal{D}|\vec{w}) \\ &= \operatorname{argmin} (-\log p(\mathcal{D}|\vec{w})) \\ &= \operatorname{argmin} (-\log L)\end{aligned}$$

Bayesian prior probability distribution over $\vec{w}$
derive mathematically the posterior from the Bayes rule
Eg: flip a coin 3 times and get 3 heads
max likelihood would give $w = 1$, w prob. of getting head
the prior compensate

Criticism: the prior is selected on the basis of mathematical convenience rather than reflection of believes.

Examples

We draw a sample $\vec{x} = (x^1, \ldots, x^m)$ from a Gaussian distribution and we want to learn the parameters of the Gaussian distribution from which the sample was drawn



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

Frequentist Approach

$\vec{X} = (X^1, \ldots, X^m)$ from $\mathcal{N}$, X_i independent and identically distributed.

$$p(\mathcal{D}|\vec{w}) = p(\vec{x}|\mu, \sigma^2) = \prod_{i=1}^{m} \mathcal{N}(x^i|\mu, \sigma^2) \qquad \text{likelihood function}$$

log is monotonically increasing function:

- it transforms $\prod$ in $\sum$
- it saves us from numerical issues with small numbers

max likelihood

$$\max \log p(\vec{x}|\mu, \sigma^2) = \log \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x^i-\mu)^2}{2\sigma^2}\right\}$$
$$= -\frac{1}{2\sigma^2}\sum_i (x^i-\mu)^2 - \frac{m}{2}\log\sigma^2 - \frac{m}{2}\log(2\pi)$$

$$\max_\mu \Longrightarrow \mu_{ML} = \frac{1}{m}\sum_{i=1}^m x_i$$

$$\max_{\sigma^2} \Longrightarrow \sigma^2_{ML} = \frac{1}{m}\sum_{i=1}^m (x_i-\mu)^2$$

(if you do not remember derivatives try http://www.wolframalpha.com)

Bayesian approach

$\vec{X} = (X^1, \ldots, X^m)$ from $\mathcal{N}$, X_i independent and identically distributed
Let's assume σ^2 known

likelihood function

$$p(\mathcal{D}|\vec{w}) = p(\vec{x}|\mu) = \prod_{i=1}^{m} \mathcal{N}(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left\{-\frac{\sum_i (x^i - \mu)^2}{2\sigma^2}\right\}$$

we choose for the prior a conjugate distribution: the posterior is again a distribution of the same form.
a Gaussian distribution has this property, hence:

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2) \qquad \text{prior distribution}$$

posterior:

$$p(\mu|\vec{X}) \propto p(\vec{X}|\mu)p(\mu)$$

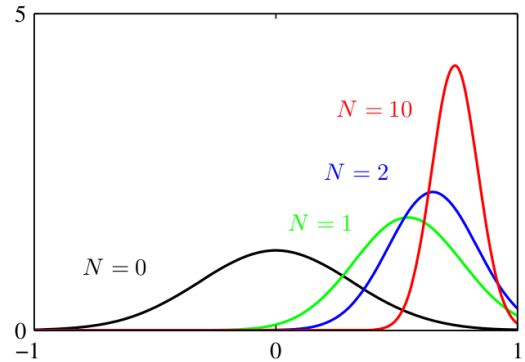
After some mathematical manipulations the posterior can be shown to be:

$$p(\mu|\vec{X}) = \mathcal{N}(\mu|\mu_m, \sigma^2)$$

$$\mu_m = \frac{\sigma^2}{m\sigma_0^2 + \sigma^2}\mu_0 + \frac{m\sigma_0^2}{m\sigma_0^2 + \sigma^2}\frac{1}{m}\sum_i x_i; \qquad \frac{1}{\sigma_m^2} = \frac{1}{\sigma_0^2} + \frac{m}{\sigma^2}$$

$$\mu_m = \frac{\sigma^2}{m\sigma_0^2 + \sigma^2}\mu_0 + \frac{m\sigma_0^2}{m\sigma_0^2 + \sigma^2}\frac{1}{m}\sum_i x_i$$

$$\frac{1}{\sigma_m^2} = \frac{1}{\sigma_0^2} + \frac{m}{\sigma^2}$$



- for $m = 0$, μ reduces to the prior
- for $m \to \infty$, μ reduces to the max likelihood solution
- the variance is more conveniently expressed in form of precision. Precisions are additive
- for $m = 0$, precision reduces to the prior
- for $m \to \infty$, variance becomes increasingly peaked around the max likelihood solution.
- hence the max likelihood solution is recovered by the Bayesian formalism in the limit of an infinite number of observations.

We show why the least square loss function
is a reasonable function for curve fitting
by looking at the problem from the probabilistic perspective.

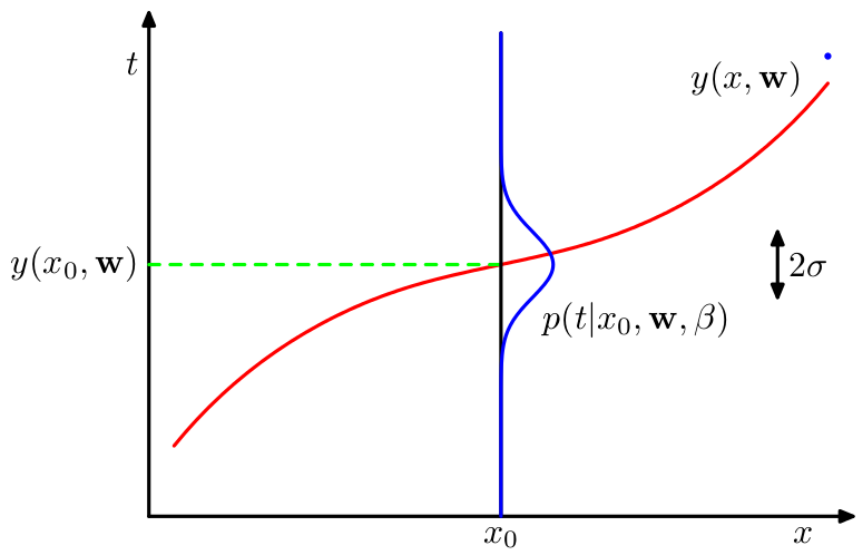
Frequentist Approach to Linear Models

$\mathbf{x} = (\vec{x}^1, \ldots, \vec{x}^m)^T$ input values

$\vec{y} = (y^1, \ldots, y^m)^T$ target values

We want to predict $\hat{y}$ for $\vec{x}$. We can express our uncertainty on $\hat{y}$ using a probability distribution. We assume:

$$p(y \mid \vec{x}, \vec{\theta}, \sigma^2) = \mathcal{N}(y \mid h(\vec{x}, \vec{\theta}), \sigma^2)$$



We search unknwons $\vec{\theta}, \sigma^2$ given training set $(\vec{x}, \vec{y})$. Data drawn independently from identical distributions

$$p(y \mid \vec{x}, \vec{\theta}, \sigma^2) = \prod_{i=1}^{m} \mathcal{N}(y^i \mid h(\vec{x}^i, \theta), \sigma^2)$$

$$\log p(y \mid \vec{x}, \vec{\theta}, \sigma^2) = -\frac{1}{2\sigma^2} \sum_i [h(\vec{x}^i, \vec{\theta}) - y^i]^2 - \frac{m}{2} \log \sigma^2 - \frac{m}{2} \log(2\pi)$$

$$\max_{\vec{\theta}} \Longrightarrow \max_{\vec{\theta}} \frac{1}{2\sigma^2} \sum_{i=1}^{m} [h(\vec{x}^i, \vec{\theta}) - y^i]^2 \qquad \text{error function}$$

$$\max_{\sigma^2} \Longrightarrow \sigma^2_{ML} = \frac{1}{m} \sum_{i=1}^{m} [h(\vec{x}^i, \vec{\theta}) - y^i]^2$$

Prediction on a new value of $\vec{x}$ uses the predictive distribution with the parameters above

$$p(y \mid \vec{x}, \vec{\theta}_{ML}, \sigma^2_{ML}) = \mathcal{N}(y \mid h(\vec{x}, \vec{\theta}_{ML}), \sigma^2_{ML})$$

it returns the expected value $E_y[p(y|\vec{w})]$

Bayesian Approach to Linear Models

Let's introduce a prior distribution over the parameters $\vec{\theta}$. For simplicity ($\alpha = 1/\sigma^2$, precision)

$$p(\vec{\theta} \mid \alpha) = \mathcal{N}(\vec{\theta} \mid \mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^2 \exp\left\{-\frac{\alpha}{2}\vec{\theta}^T\vec{\theta}\right\}$$

by Bayes' theorem:

$$p(\vec{\theta} \mid y, \vec{x}, \alpha, \sigma^2) \propto p(y \mid \vec{x}, \vec{\theta}, \sigma^2)p(\vec{\theta} \mid \alpha)$$

we find $\vec{\theta}$ by finding most probable value given data (max posterior)

$$\min\left\{-\log p(y \mid \vec{x}, \vec{\theta}, \sigma^2) - \log p(\vec{\theta} \mid \alpha)\right\}$$

$$\min\{-\frac{1}{2\sigma^2}\sum_i [h(\vec{x}^i, \vec{\theta}) - y^i]^2 - \frac{m}{2}\log\sigma^2 - \frac{m}{2}\log(2\pi)$$

$$+ (p+1)\log(\alpha 2\pi) + \frac{\alpha}{2}\vec{\theta}^T\vec{\theta}\}$$

$$\min\left\{-\frac{1}{2\sigma^2}\sum_i [h(\vec{x}^i, \vec{\theta}) - y^i]^2 + \frac{\alpha}{2}\vec{\theta}^T\vec{\theta}\right\}$$ quadratic loss + regularization

However we are still making point estimate.
A full Bayesian approach uses consistently only sum and product rule of probabilities:

$(\vec{y}, \mathbf{x})$ training data + new point $\vec{x}$.
We want to predict $\hat{y}$ for $\vec{x}$.
Hence we are interested in the predictive distribution $p(y \mid \mathcal{D})$, ie $p(y \mid \vec{x}, \vec{y}, \mathbf{x})$

$$p(y \mid \vec{x}, \vec{y}, \mathbf{x}, \alpha, \beta) = \int p(y \mid \vec{x}, \vec{\theta}) p(\vec{\theta} \mid y, \vec{x}, \alpha, \beta) d\vec{\theta}$$

everything can be derived analytically and is of the form:

$$p(y \mid \vec{x}, \vec{y}, \mathbf{x}, \alpha, \beta) = \mathcal{N}(y \mid m(\vec{x}), s^2(\vec{x}))$$