Artificial Intelligence (EI06024001)

2: Introduction to ML

2.1 Probability theory

A key concept in the field of pattern recognition is that of uncertainty. It arises both through noise on measurements, as well as through the finite size of data sets. Probability theory provides a consistent framework for the quantification and manipulation of uncertainty and forms one of the central foundations for pattern recognition.

2.2. Decision theory

Combined with probability theory, it can allow us to make optimal decisions in situations involving uncertainty.

2.3. Information theory

A key measure in information theory is "entropy". Entropy quantifies the amount of uncertainty involved in the value of a random variable or the outcome of a random process.

2.1 Probability Theory

Two fundamental rules of probability theory

sum rule $$p(X) = \sum_Y p(X, Y)$$

product rule $$p(X, Y) = p(Y|X)p(X).$$

Bayes' theorem

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

with $$p(X) = \sum_Y p(X|Y)p(Y).$$

2.1 Probability Theory

Bayes' rule (theorem)

- H: hypothesis　가정
- e: event　사건

Likelihood

How probable is the evidence given that hypothesis is true?

가정이 사실일 때, 사건에 대한 확률

Prior

How probable was our hypothesis before observing the evidence?

가정에 대한 사전확률

$$P(H|e) = \frac{P(e|H) \cdot P(H)}{P(e)}$$

Posterior

How probable is our hypothesis given the observed evidence? (Not directly computed)

사건이 관측되었을 때, 가정에 대한 확률

Marginal

How probable is the evidence under all possible hypotheses?

$$P(e) = \sum_i P(e|H_i)P(H_i)$$

모든 가정에서, 사건에 대한 확률

2.1 Probability Theory

Bayes' rule (theorem)

- Scenario:
  - H(hypothesis): "I have a cold"
  - e(event): "I have a runny nose"

$$P(H|e) = \frac{P(e|H) \cdot P(H)}{P(e)}$$

  - Posterior $P(H|e)$: the probability that I have a cold, given that I have a runny nose
  - Likelihood $P(e|H)$: the probability of having a runny nose when I have a cold
  - Prior $P(H)$: the probability of having a cold, without knowing what my symptoms are
  - Marginal $P(e)$: the probability of having a runny nose, whatever the cause may be

2.1 Probability Theory

Bayes' rule (theorem)

- H: hypothesis   가정   → w: model parameters
- e: event   사건   → D: data

$$P(H|e) = \frac{P(e|H) \cdot P(H)}{P(e)}$$

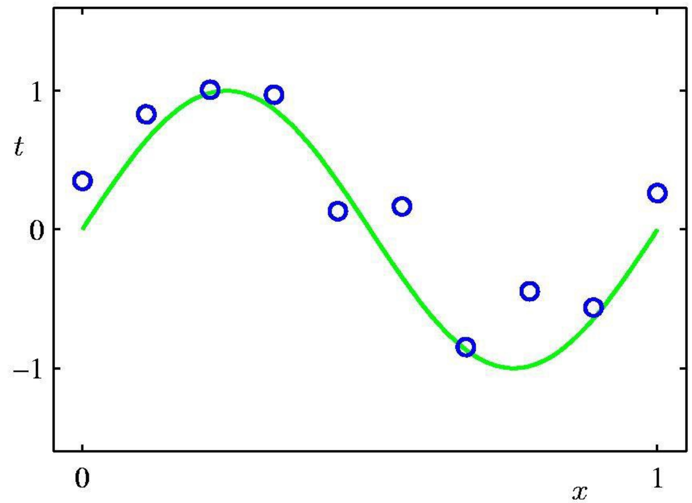
↓

$$P(w|D) = \frac{P(D|w) \cdot P(w)}{P(D)}$$

2.1 Probability Theory

Bayesian probabilities:

In polynomial curve fitting problem,



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

$y \longrightarrow y^*$: find y that is closed to y^*

$w \longrightarrow w^*$: find $\mathbf{w}$ that is closed to $\mathbf{w}^*$

2.1 Probability Theory

Bayesian probabilities:

In polynomial curve fitting problem,

$$P(w|D) = \frac{P(D|w) \cdot P(w)}{P(D)}$$

- Prior $P(w)$: assumption on w before observing D
- Likelihood $P(D|w)$: how probable the observed data set D is for different settings of the parameter w
- Marginal $P(D)$: distribution of data set D
- Posterior $P(w|D)$: certainty in w after we have observed D

2.1 Probability Theory

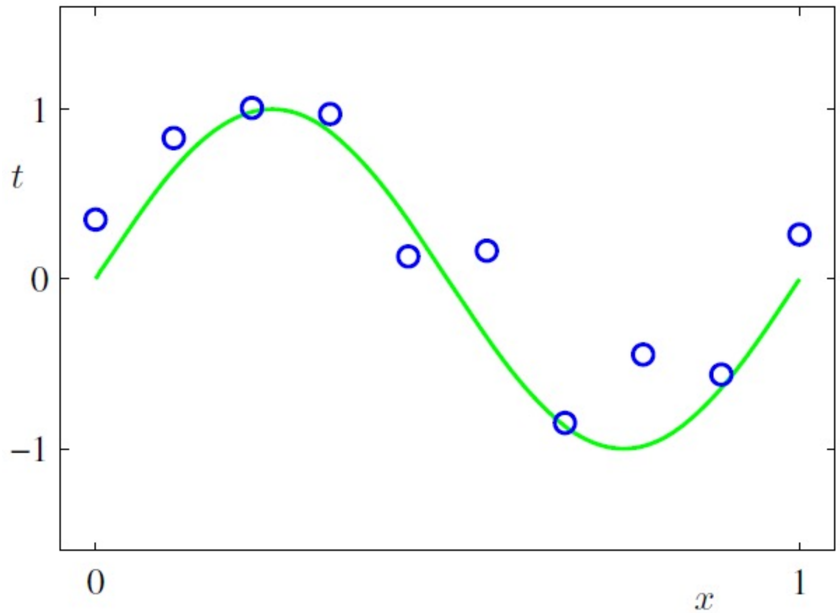
Bayesian probabilities:

Frequentist estimator: maximum likelihood (MLE);

Bayesian estimator: maximum a posteriori (MAP);

$$\mathbf{P}(\boldsymbol{w}|\boldsymbol{D}) = \frac{\mathbf{P}(\boldsymbol{D}|\boldsymbol{w}) \cdot P(w)}{P(D)}$$

Example: Polynomial Curve Fitting Problem



Training set: $\mathbf{x} \equiv (x_1, \dots, x_N)^{\mathrm{T}}$ Input value

$\mathbf{t} \equiv (t_1, \dots, t_N)^{\mathrm{T}}$ Target value

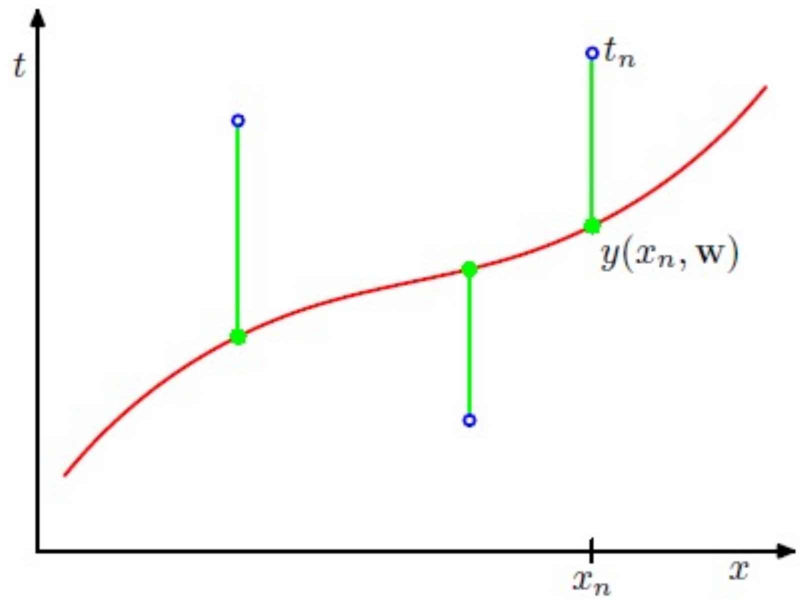
Goal: Given a new x, make the prediction of t

Example: Polynomial Curve Fitting Problem

Solution 1: Error minimization

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Minimize $$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

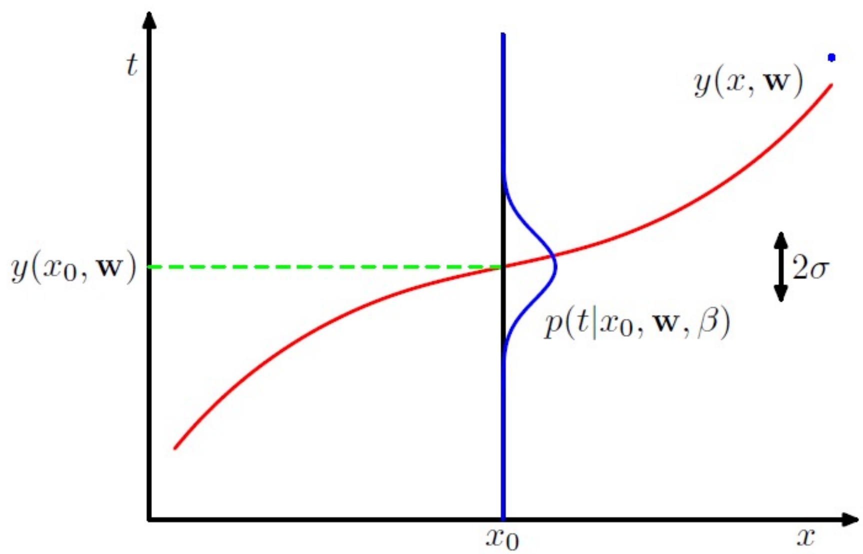


Example: Polynomial Curve Fitting Problem

Solution 2: A probabilistic perspective

Assume that, given the value of x, the corresponding value of t has a Gaussian distribution with a mean equal to the value $y(x, \mathbf{w})$.

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



Example: Polynomial Curve Fitting Problem

Solution 3: A Bayesian treatment

A more Bayesian approach: introduce a *prior distribution* over the polynomial coefficients $\mathbf{w}$.

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

Assume that the parameters α and β are fixed and known.

Maximize $\quad p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) \cdot p(\mathbf{w}|\alpha)$

$$\Downarrow$$

Minimize $\quad \dfrac{\beta}{2}\displaystyle\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$

Example: Polynomial Curve Fitting Problem

Solution 3: A Bayesian treatment

Proof:

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

$$\prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1}) \cdot \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

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

$$\prod_{n=1}^{N} \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\left\{-\frac{\beta}{2}\left(t_n - y(x_n, \mathbf{w})\right)^2\right\} \cdot \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

$$\left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left\{-\frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - y(x_n, \mathbf{w})\right)^2\right\} \cdot \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

$$\left(\frac{\beta}{2\pi}\right)^{N/2} \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - y(x_n, \mathbf{w})\right)^2 - \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

Example: Polynomial Curve Fitting Problem

Solution 3: A Bayesian treatment

Proof:

$$p(\mathbf{w}|\mathbf{x},\mathbf{t},\alpha,\beta) \propto \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - y(x_n,\mathbf{w})\right)^2 - \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

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

$\downarrow$

Maximize $\exp\left\{-\frac{\beta}{2}\sum_{n=1}^{N}\left(t_n - y(x_n,\mathbf{w})\right)^2 - \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$

$\downarrow$

Minimize $\frac{\beta}{2}\sum_{n=1}^{N}(y(x_n,\mathbf{w}) - t_n)^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$

Example: Polynomial Curve Fitting Problem

Solution 3: A Bayesian treatment

Proof:

Minimize $$\frac{\beta}{2}\sum_{n=1}^{N}(y(x_n,\mathbf{w})-t_n)^2+\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

Solution 1: Error minimization

Minimize $$E(\mathbf{w})=\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2+\frac{\lambda}{2}\|\mathbf{w}\|^2$$

Example: Polynomial Curve Fitting Problem

Solution 3: A Bayesian treatment

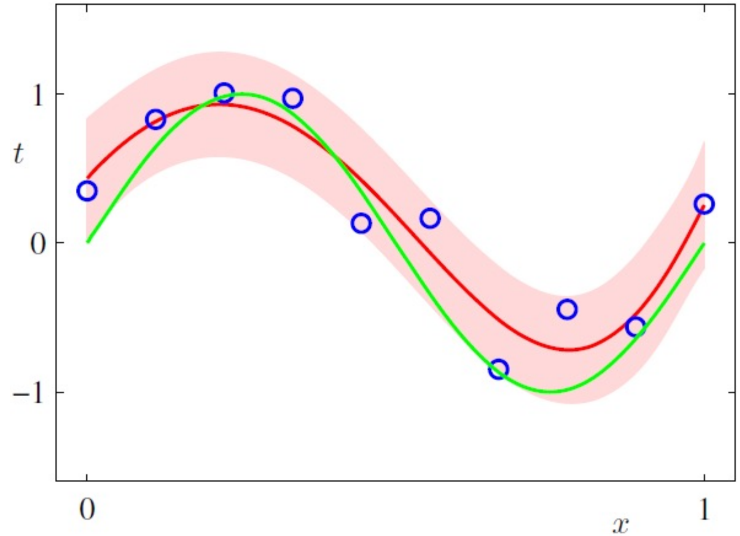
In a fully Bayesian approach, we should consistently apply the sum and product rules of probability, which requires that we integrate over all values of $\mathbf{w}$.

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) d\mathbf{w}$$

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right)$$

$$m(x) = \beta \phi(x)^{\mathrm{T}} \mathbf{S} \sum_{n=1}^{N} \phi(x_n) t_n$$

$$s^2(x) = \beta^{-1} + \phi(x)^{\mathrm{T}} \mathbf{S} \phi(x)$$



Where,

$$\phi(x_n) = (1, x_n, x_n^2, \dots, x_n^M)^{\mathrm{T}}$$

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