CC-226 Aula 07 - Estimação de Parâmetros

Carlos Henrique Q. Forster - Instituto Tecnológico de Aeronáutica

2008

Estimação de Parâmetros

- Para construir o classificador bayesiano, assumimos as distribuições de probabilidades conhecidas.
- Para as redes bayesianas e o método naïve bayes, haviamos utilizado apenas distribuições discretas.
- Agora queremos estimar uma densidade de probabilidade contínua (e eventualmente multidimensional) a partir de um conjunto de observações.
- Se conhecemos o tipo de variável aleatória com que estamos trabalhando, podemos conhecer a família de distribuições de probabilidades (uniforme, gaussiana, Rayleigh etc) e neste caso bastaria determinar um vetor de parâmetros que definem a forma da distribuição.
- Há casos em que esses parâmetros não são conhecidos e devem ser estimados a partir dos dados (adaptar um modelo paramétrico aos dados como fizemos na regressão linear).
- No problema da regressão linear, minimizamos uma função de erro que media a disparidade entre os dados e um modelo de predição. O que podemos utilizar como função de erro para avaliar se um conjunto de parâmetros modela adequadamente um conjunto de amostras?

Amostragem

$X_1, \ldots, X_N$ formam uma amostra aleatória de tamanho N se

- X_i são todos independentes, ou seja, $P(X_i, X_j) = P(X - i)P(X_j), \forall i \neq j$;
- Todo X_i possui a mesma distribuição de probabilidades.

Propriedade: distribuição da média amostral

Se os dados $X_1, \ldots, X_N$ foram amostrados de uma distribuição qualquer com valor médio μ e desvio-padrão σ:

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

Se $T_0 = X_1 + \ldots + X_n$, então:

$$E(T_0) = n\mu$$

$$Var(T_0) = n\sigma^2$$

Propriedade: teorema do limite central

Se os dados $X_1, \ldots, X_N$ foram amostrados de uma distribuição qualquer com valor médio μ e desvio-padrão σ. Se N é suficientemente grande, a média amostral $\bar{X}$ tem aproximadamente uma distribuição normal com

$$\mu_{\bar{X}} = \mu$$

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

A soma $T_0 = \sum_i X_i$ é também aproximadamente normal com

$$\mu_{T_0} = n\mu$$

$$\sigma^2_{\bar{X}} = n\sigma^2$$

Quanto maior for n melhor a aproximação.

Estimação por Máxima Verossimilhança (MLE)

- Na construção do classificador, assumimos conhecida a verossimilhança $p(x|C_i), i = 1 \ldots M$ de como o conjunto de treinamento se distribuía.
- Supomos que essa verossimilhança assume uma forma paramétrica,dependente do vetor de parâmetros θ_i desconhecido e que desejamos estimar.
- Sejam $x_1 \ldots x_N$, N observações com pdf $p(x, \theta)$.
- O conjunto $D = \{x_k\}$ é o conjunto de amostras com probabilidade conjunta $p(D, \theta)$.

- Os dados do conjunto D são independentes e identicamente distribuídos, de forma que:

$$p(D, \theta) = p(x_1, x_2, \ldots, x_N) = \prod_{k=1}^{N} p(x_k, \theta)$$

- O estimador de máxima verossimilhança $\hat{\theta}_{ML}$ é aquele θ para qual a função de verossimilhança tem valor máximo:

$$\hat{\theta}_{ML} = arg \max_{\theta} \prod_{k=1}^{N} p(x_k, \theta)$$

- Na otimização irrestrita, a condição necessária para obter o máximo é que o gradiente da função de verossimilhança em relação ao vetor θ seja nulo.

$$\frac{\partial \prod_{k=1}^{N} p(x_k, \theta)}{\partial \theta} = 0$$

- O logaritmo é crescente e monotônico, podemos definir a função log-verossimilhança como

$$\mathcal{L}(\theta) = \ln \prod_{k=1}^{N} p(x_k, \theta)$$

- Operando o gradiente do log:

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \sum_{k=1}^{N} \frac{\partial \ln p(x_k, \theta)}{\partial \theta} = \sum_{k=1}^{N} \frac{1}{p(x_k, \theta)} \frac{\partial p(x_k, \theta)}{\partial \theta} = 0$$

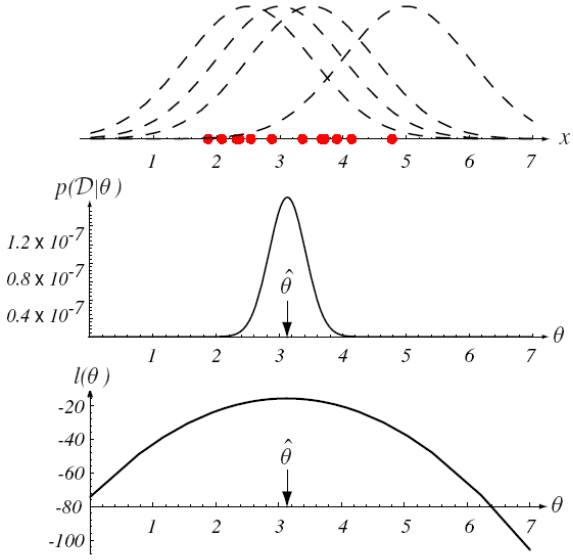


FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Propriedades da MLE

- Se θ_0 é o valor real, o estimador ML é assintoticamente não-viezado, isto é, para amostras grandes ($n \to \infty$) em média converge para o valor real.

$$\lim_{n\to\infty} \mathcal{E}[\hat{\theta}_{ML}] = \theta_0$$

- Em geral, o estimador ML pode ser viezado para amostras pequenas (lembrar do σ^2).

- Porém, é assintoticamente consistente: a probabilidade de estar arbitrariamente próximo do valor correto é alta quanto maior for N.

$$\lim_{n\to\infty} Pr[\|\hat{\theta}_{ML} - \theta_0\| < \varepsilon] = 1$$

- O estimador ML converge na média dos quadrados, isto é, para amostras grandes, a variância do estimador ML tende a zero.

$$\lim_{n\to\infty} \mathcal{E}[\|\hat{\theta}_{ML} - \theta_0\|^2] = 0$$

- O estimador de ML tende a ser o estimador de mínima variância e é aproximadamente não-tendencioso.
- A pdf do estimador ML tende a uma gaussiana.
- Princípio da invariância: Se $\hat{\theta}$ é um estimador de máxima verossimilhança para o vetor de estatísticas θ, o estimador de máxima verossimilhança das estatísticas $h(\theta)$ é $h(\hat{\theta})$ qualquer que seja h.

Exemplo

Caso Gaussiano com μ desconhecido.

Supomos Σ conhecido.

$$\ln p(x_k|\mu) = -\frac{1}{2}\ln\left[(2\pi)^d|\Sigma|\right] - \frac{1}{2}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)$$

$$\nabla_\mu \ln p(x_k|\mu) = \Sigma^{-1}(x_k-\mu)$$

$$\sum_{k=1}^{n}\Sigma^{-1}(x_k-\mu) = 0$$

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

O estimador ML da média da distribuição é a média amostral.

Exemplo

Caso Gausiano com μ e Σ desconhecidos.

Consideramos o caso unidimensional com $\theta_1 = \mu$ e $\theta_2 = \sigma^2$

$$\ln p(x_k|\theta) = \frac{-1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k-\theta_1)^2$$

Com derivadas

$$\nabla_\theta \mathcal{L} = \nabla_\theta \ln p(x_k|\theta) = \begin{bmatrix} \frac{1}{\theta_2}(x_k-\theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2{\theta_2}^2} \end{bmatrix}$$

Igualando a zero,

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0$$

$$-\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^{2^2}}{\hat{\theta}_2} = 0$$

Assim,

$$\theta_1 = \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$\theta_2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \hat{\mu})^2$$

No caso multivariado:

$$\hat{\mu} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{k=1}}^{\mathbf{n}} \mathbf{x_k}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n} (x_k - \hat{\mu})(x_k - \hat{\mu})^T$$

Notar que o estimador ML para a variância é tendencioso (viezado).

$$\mathcal{E}\left[\frac{1}{n}\sum_{i=1}^{n} (x_k - \bar{x})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

Um estimador não tendencioso seria:

$$C = \frac{1}{n-1}\sum_{k=1}^{n} (x_k - \hat{\mu})(x_k - \hat{\mu})^T$$

Estimador Máxima a posteriori (MAP)

- Considerando θ como uma variável aleatória, podemos acrescentar informação sobre o parâmetro $theta$ baseado em conhecimento do problema.
- Essa informação é codificada como uma distribuição de probabilidade *a priori* $p(\theta)$.
- Seja $D = x_1, \ldots, x_N$ o conjunto de amostra de tamanho N, obtido de forma i.i.d.
- O estimador para θ que procuramos é aquele que minimiza a probabilidade *a posteriori* $p(\theta|D)$.

- Expandindo a regra de Bayes:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- O estimador MAP é aquele $\hat{\theta}$ para o qual:

$$\frac{\partial}{\partial\theta}p(\theta|D) = 0$$

ou

$$\frac{\partial}{\partial\theta}\left(p(\theta)p(D|\theta)\right) = 0$$

- Dependendo de $p(\theta)$ a priori, o resultado diferirá do estimador ML.
- Podemos encontrar uma distribuição a priori que não seja informativa, de forma a obter um estimador ML. Em geral, isto é simples se o espaço de parâmetros for limitado. No caso de espaço não limitado, a regra que a integral da probabilidade é igual a 1 é ignorada, de forma que a distribuição a priori é uma distribuição imprópria.
- Se a verossimilhança pertence a uma família de distribuições de probabilidade, podemos encontrar uma família de distribuições de probabilidades que seja conjugada daquela, ou seja, a forma da pdf a posteriori é da mesma família da pdf a priori.
- No caso da distribuição de Bernoulli, a pdf conjugada é uma distribuição beta, no caso da Gaussiana, a pdf conjugada para a média é também uma Gaussiana. Para a variância da Gaussiana, a priori conjugada é uma distribuição de Wishart.

Sumário

- Verossimilhança é a função de θ dada por $p(D|\theta)$.
- Escolher o $theta$ que maximiza a verossimilhança corresponde a escolher o conjunto de parâmetros que melhor explica os dados D.
- MLE (maximum likelihood estimator) estima pelo máximo $\mathcal{L}(\theta)$.
- MAP (maximum a posteriori) estima pelo máximo $\mathcal{L}(\theta) + \ln p(\theta)$.

Estimação Bayesiana

- Estimação bayesiana não considera estimação pontual, mas considera θ como uma variável aleatória, cuja probabilidade a posteriori será convertida numa distribuição a posteriori baseado num conjunto de dados.

- $p(x)$ é a probabilidade do fenômeno que estamos interessados a estimar, assumimos que tem uma forma paramétrica baseada num vetor de parãmetros θ.
- $p(x)$ é desconhecido, mas $p(x|\theta)$ é completamente determinado (a partir da suposição acima).
- $p(\theta)$ é a distribuição de probabilidade a priori sobre os parâmetros, isto é, as informações que temos sobre os parâmetros antes de realizar as observações.
- $p(\theta|D)$ é a informação posterior, esperamos que tenha um pico em torno do valor verdadeiro de θ.
- $p(x|D)$ é uma aproximação para $p(x)$, baseada nos dados D.

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

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

visto que

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

Exemplo

Seja σ^2 conhecido e $D = \{x_k\}$ um conjunto de amostra i.i.d.

Estimar a probabilidade para a média μ.

Temos também a informação adicional de que a média segue uma distribuição normal com média μ_0 e variância σ_0^2 conhecidas.

$$p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

$$p(D|\mu) \sim \mathcal{N}(\mu, \sigma^2)$$

Queremos $p(\mu|D)$

Aplicando a regra de Bayes,

$$p(\mu|D) = \frac{p(D|\mu)p(\mu)}{\int p(D|\mu)p(\mu)d\mu}$$

$$= \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu)$$

O produto e a soma das normais resulta numa distribuição normal em μ:

$$p(\mu|D) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n}\right)^2\right]$$

Com

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right) \hat{\mu} + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0$$

$$\sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}$$

Em seguida, obtemos:

$$p(x|D) = \int p(x|\mu) p(\mu|D) d\mu$$

Obtemos que

$$p(x|D) \sim \mathcal{N}(\mu_n, \sigma^2 + \sigma_n^2)$$

Junto com probabilidades a priori sobre classes $P(C_j)$, podemos obter as posteriores $P(C_j|x)$ e construir um classificador. Basta para cada classe C_j utilizar o conjunto de dados correspondente D_j.

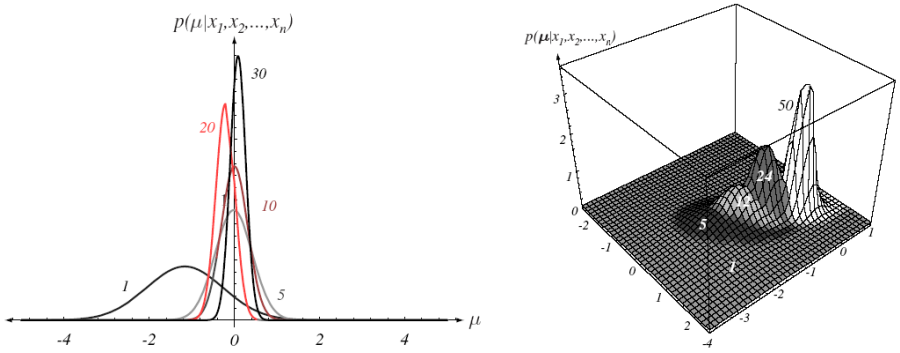


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Forma Geral - Estimação Bayesiana

- $p(x|\theta)$ conhecida, θ desconhecida
- $p(\theta)$ conhecida
- Para um conjunto de amostra $D = x_k, k = 1 \ldots N$
  - Integrar por todo o espaço de parâmetros (usando, por exemplo, Monte Carlo - algoritmo de Gibbs)

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

– basta obter $p(\theta|D)$ utilizando

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

– usando a independência das amostras:

$$p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$$

- O método pode ser feito *on-line* utilizando a forma recursiva do aprendizado Bayesiano.

$$p(\theta|D^0) = p(\theta)$$

$$p(\theta|D^n) = \frac{p(x_n|\theta)p(\theta|D^{n-1})}{\int p(x_n|\theta)p(\theta|D^{n-1})d\theta}$$

Exemplo: População com Distribuição Uniforme

$$p(x|\theta) \sim U(0,\theta) = \begin{cases} \frac{1}{\theta}, & 0 \leq x \leq \theta \\ 0, & \text{caso contrário} \end{cases}$$

Sabemos a priori que $0 < \theta \leq 10$
Dados $D = \{4, 7, 2, 8\}$.
Para o conjunto vazio de dados:

$$p(\theta|D^0) = p(\theta) = U(0, 10)$$

Recebendo a primeira observação da amostra:

$$x_1 = 4 \rightarrow p(\theta|D^1) \propto p(x|\theta)p(\theta|D^0)$$

$$= \begin{cases} \frac{1}{\theta}, & 4 \leq \theta \leq 10 \\ 0, & \text{caso contrário} \end{cases}$$

Recebendo a segunda observação da amostra:

$$x_2 = 7 \rightarrow p(\theta|D^2) \propto p(x|\theta)p(\theta|D^1)$$

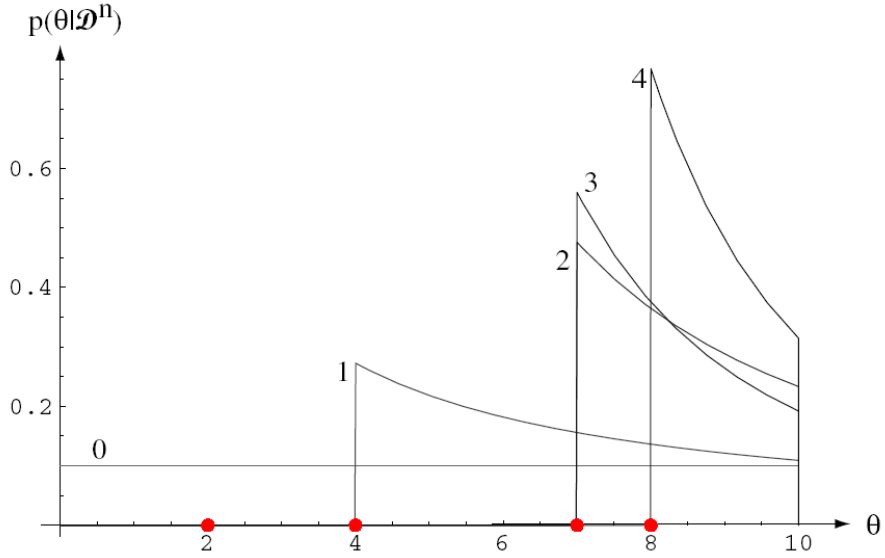
$$= \begin{cases} \frac{1}{\theta^2}, & 7 \leq \theta \leq 10 \\ 0, & \text{caso contrário} \end{cases}$$

Para as demais observações:

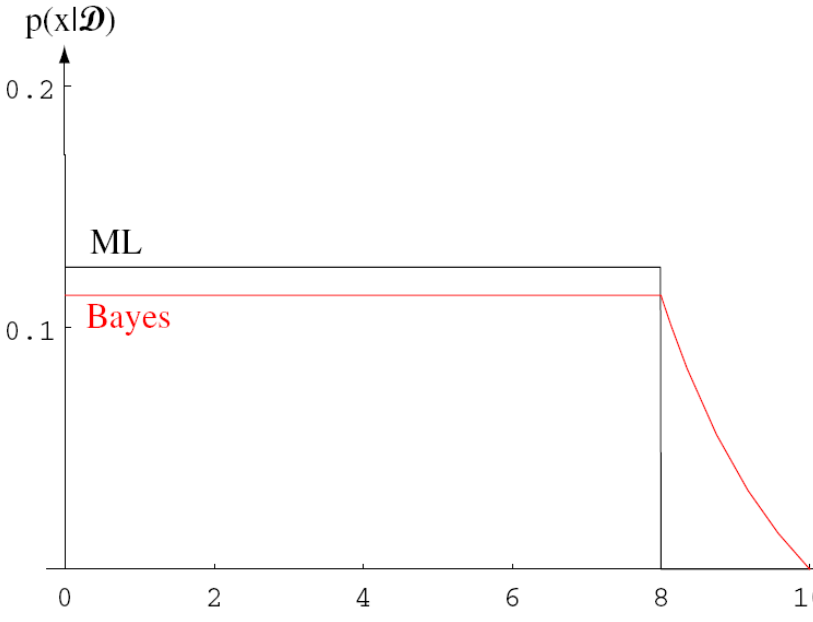
$$p(\theta|D^n) \propto \frac{1}{\theta^n}$$

para

$$\max_x [D^n] \leq \theta \leq 10$$



Solução MLE é pontual: $\hat{\theta} = 8$ então $p(x|D) = p(x|\hat{\theta})$.



Solução bayesiana é uma distribuição:

Estatísticas Suficientes

- Uma estatística é qualquer valor calculado a partir de uma amostra dada.
- Estatísticas suficientes consistem no vetor de estatísticas que é suficiente para determinar um conjunto de estatísticas da distribuição.
- Estimar a densidade (ou seus parâmetros) pode consistir basicamente em estimar esse vetor de estatísticas.

Exemplo de MLE: distribuição de Bernoulli

A probabilidade de x ser 0 ou 1:

$$P(x) = p^x(1-p)^{1-x}$$

Onde p é o parâmetro da distribuição.
Estimar p a partir de um conjunto D de amostras.

$$\mathcal{L}(p|D) = \log \prod_{k=1}^{N} p(x_k)(1-p)^{(1-x_k)}$$

$$= \sum_{k=1}^{N} x_k \log p + \left(N - \sum_k x_k \right) \log (1-p)$$

Derivando em p e igualando a zero, obtemos:

$$\hat{p} = \frac{\sum_k x_k}{N}$$

Regressão e MLE

No modelo da regressão:

$$r = f(x) + \varepsilon$$

A função $g(x|\theta)$ aproxima $f(x)$, uma vez determinados os θ.
Supomos que $\varepsilon \sim \mathcal{N}(g(x|\theta), \sigma^2)$.

$$p(r|x) \sim \mathcal{N}(g(x|\theta), \sigma^2)$$

Os pares $D = \{(x_k, r_k)\}$ foram obtidos da densidade de probabilidade $p(x, r)$.
Abrindo em probabilidade condicional e probabilidade a priori, temos:

$$p(x, r) = p(r|x)p(x)$$

Vamos estimar os parâmetros θ dados D. A função de verossimilhança é:

$$\mathcal{L}(\theta|D) = \log \prod_{k=1}^{N} p(x_k, r_k)$$

$$= \log \prod_{k=1}^{N} p(r_k|x_k) + \log \prod_{k=1}^{N} p(x_k)$$

A segunda parcela é constante e, portanto, ignorada.

Considerando o modelo de distribuição normal:

$$= \log \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{[r_k - g(x_k|\theta)]^2}{2\sigma^2}\right]$$

$$= -N \log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2} \sum_{k=1}^{N} [r_k - g_k|\theta]^2$$

Assim, para obter o estimador de máxima verossimilhança para os parâmetros da regressão, basta minimizar:

$$E(\theta|D) = \frac{1}{2} \sum_{k=1}^{N} [r_k - g(x_k|\theta)]^2$$

Ou seja, é a própria estimação pelo critério dos mínimos quadrados.

No caso da regressão polinomial, se considerarmos uma probabilidade a priori sobre os valores de θ, com média zero e dada variância, limitamos o crescimento dos parâmetros do polinômio. Assim, uma dist. a priori da forma

$$p(\theta) = \exp\left(-\frac{\alpha}{2}\theta^T\theta\right)$$

torna-se uma parcela na função a ser minimizada:

$$+\frac{\alpha}{2}\theta^T\theta$$

que é a parcela de regularização e α é a constante de regularização (o λ da aula sobre regressão).