Patt. Rec. and Mach. Learning
Ch. 1: Introduction

Radu Horaud & Pierre Mahé

September 28, 2007

Chapter content

- Goals, terminology, scope of the book;
- 1.1 Example: Polynomial curve fitting;
- 1.2 Probability theory;
- 1.3 Model selection;
- 1.4 The curse of dimensionality;
- 1.5 Decision theory;
- 1.6 Information theory.

Goals

Pattern Recognition: automatic discovery of regularities in data and the use of these regularities to take actions – *classifying the data into different categories.* Example: handwritten recognition. Input: a vector $\boldsymbol{x}$ of pixel values. Output: A digit from 0 to 9.

Machine learning: a large set of input vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, or a *training set* is used to tune the parameters of an adaptive model. The *category* of an input vector is expressed using a *target vector* $\boldsymbol{t}$. The result of a machine learning algorithm: $\boldsymbol{y}(\boldsymbol{x})$ where the output $\boldsymbol{y}$ is encoded as the target vectors.

Terminology

- *training* or *learning* phase: determine $\boldsymbol{y}(\boldsymbol{x})$ on the basis of the *training data*.
- *test set*, *generalization*,
- *supervised learning* (input/target vectors in the training data),
- *classification* (discrete categories) or *regression* (continuous variables),
- *unsupervised learning* (no target vectors in the training data) also called *clustering*, or *density estimation*.
- *reinforcement learning*, *credit assignment*, *exploration*, *exploitation*.

1.1 Polynomial curve fitting

- Training set: $\mathbf{x} \equiv (x_1, \ldots, x_N)$ AND $\mathbf{t} \equiv (t_1, \ldots, t_N)$
- Goal: predict the target $\hat{t}$ for some new input $\hat{x}$
- *Probability theory* allows to express the uncertainty of the target.
- *Decision theory* allows to make optimal predictions.

- Minimize:

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2$$

- The case of a polynomial function linear in $\boldsymbol{w}$:

$$y(x_n, \boldsymbol{w}) = \sum_{j=0}^{M} w_j x^j$$

- *Model selection*: choosing M.
- *Regularization* (adding a penalty term):

$$\tilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$$

- This can be expressed in the Bayesian framework using *maximum likelihood*.

1.2 Probability theory (discrete random variables)

- Sum rule:

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

- Product rule:

$$p(X, Y) = p(X|Y)p(Y) = p(Y|X)p(X)$$

- Bayes:

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

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{normalization}}$$

1.2.1 Probability densities (continuous random variables)

- Probability that x lies in an interval:

$$p(x \in (a,b)) = \int_a^b p(x)dx$$

- $p(x)$ is called the *probability density* over x.
- $p(x) \geq 1$, $p(x \in (-\infty, \infty)) = 1$
- nonlinear change of variable $x = g(y)$:

$$p_y(y) = p_x(x) \left| \frac{dx}{dy} \right|$$

- *cumulative distribution function*: $P(z) = p(x \in (-\infty, z))$
- sum and product rules extend to probability densities.

1.2.2 Expectations and covariances

- Expectation: the average value of some function $f(x)$ under a probability distribution $p(x)$;
- discrete case: $E[f] = \sum_x p(x)f(x)$
- continuous case: $E[f] = \int p(x)f(x)dx$
- N points *drawn from the prob. distribution or prob. density*, expectation can be approximated by:

$$E[f] \approx \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

- functions of several variables: $E_x[f] = \sum_x p(x)f(x, y)$ (MODIFIED)
- conditional expectation: $E_x[f|y] = \sum_x p(x|y)f(x)$

- *Variance* of $f(x)$: a measure of the variations of $f(x)$ around $E[f]$.
- $var[f] = E[f^2] - E[f]^2$
- $var[x] = E[x^2] - E[x]^2$
- *Covariance* for two random variables: $cov[x, y] = E_{x,y}[xy] - E[x]E[y]$
- Two vectors of random variables: $cov[\boldsymbol{x}, \boldsymbol{y}] = E_{x,y}[\boldsymbol{x}\boldsymbol{y}^\top] - E[\boldsymbol{x}]E[\boldsymbol{y}^\top]$
- *ADDITIONAL FORMULA*:

$$E_{x,y}[f(x, y)] = \sum_x \sum_y p(x, y) f(x, y)$$

1.2.3 Bayesian probabilities

- frequentist versus Bayesian interpretation of probabilities;
- frequentist estimator: *maximum likelihood* (MLE or ML);
- Bayesian estimator: MLE and *maximum a posteriori* (MAP);
- back to curve fitting: $\mathcal{D} = \{t_1, \dots, t_N\}$ is a set of N observations of N random variables, and $\boldsymbol{w}$ is the vector of unknown parameters.

  매개변수 w에 대한 함수

  관측데이터 집합 D를 바탕으로 계산
- Bayes theorem writes in this case: $p(\boldsymbol{w}|\mathcal{D}) = \dfrac{p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})}{p(\mathcal{D})}$
- posterior $\propto$ likelihood $\times$ prior (all these quantities are parameterized by $\boldsymbol{w}$) 전부 w에 대한 함수이다.
- $p(\mathcal{D}|\boldsymbol{w})$ is the *likelihood function* and denotes how probable is the observed data set for various values of $\boldsymbol{w}$. It is not a probability distribution over $\boldsymbol{w}$.
- The denominator: P(D)는 이렇게 쓸수있다.

$$p(\mathcal{D}) = \int \dots \int p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$$

1.2.4 The Gaussian distribution

- The Gaussian distribution of a single real-valued variable x:

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

- in D dimensions: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) : \mathbb{R}^D \rightarrow \mathbb{R}$
- $E[x] = \mu$, $\text{var}[x] = \sigma^2$
- $\mathbf{x} = (x_1, \ldots, x_N)$ is a set of N observations of the SAME scalar variable x
- Assume that this data set is *independent and identically distributed*:

$$p(x_1, \ldots, x_N|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$

- $\max p$ is equivalent to $\max \ln(p)$ or $\min(-\ln(p))$
- $\ln p(x_1, \ldots, x_N|\mu,\sigma^2) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \ldots$
- maximum likelihood solution: μ_{ML} and σ^2_{ML}
- MLE underestimates the variance: bias

1.2.5 Curve fitting re-visited

- training data: $\mathbf{x} = (x_1, \ldots, x_N)$ and $\mathbf{t} = (t_1, \ldots, t_N)$
- it is assumed that t is Gaussian: $p(t|x, \boldsymbol{w}, \beta) = (t|y(x, \boldsymbol{w}), \beta^{-1})$
- recall that: $y(x, \boldsymbol{w}) = w_0 + w_1 x + \ldots + w_M x^M$
- (joint) likelihood function: $p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \boldsymbol{w}), \beta^{-1})$
- log-likelihood:

$$\ln p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{t_n - y(x_n, \boldsymbol{w})\}^2 + \frac{N}{2} \ln \beta - \ldots$$

- $\beta = \frac{1}{\sigma^2}$ is called the precision.
- The ML solution can be used as a *predictive distribution*: $p(t|x, \boldsymbol{w}_{ML}, \beta_{ML}) = (t|y(x, \boldsymbol{w}_{ML}), \beta_{ML}^{-1})$

Introducing a prior distribution

- The polynomial coefficients are treated as random variables with a Gaussian distribution taken over a vector of dimension $M+1$:

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

- from Bayes we get the posterior probability: $p(\boldsymbol{w}|\mathbf{x},\mathbf{t},\alpha,\beta) \propto p(\mathbf{t}|\mathbf{x},\boldsymbol{w},\beta)p(\boldsymbol{w}|\alpha)$
- maximum posterior or MAP. We take the negative logarithm, we throw out constant terms and we get:

$$\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - y(x_n,\boldsymbol{w})\}^2 + \frac{\alpha}{2}\boldsymbol{w}^\top \boldsymbol{w}$$

1.2.6 Bayesian curve fitting

- Apply the correct Bayes formula:

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

- Section 3.3: the posterior distribution is a Gaussian and can be evaluated analytically.
- the sum and product rules can be used to compute the predictive distribution:

$$p(t|x,\mathbf{x},\mathbf{t}) = \int p(t|x,\boldsymbol{w})p(\boldsymbol{w}|\mathbf{x},\mathbf{t})d\boldsymbol{w} = \mathcal{N}(t, m(x), s^2(x))$$

1.3 Model selection

- which is the optimal order of the polynomial that gives the best generalization?
- train a range of models and test them on an independent *validation set*
- cross-validation: use a subset for training and the whole set for assessing the performance
- Akaike information criterion: $\ln p(\mathcal{D}|\boldsymbol{w}_{ML}) - M$
- Bayesian information criterion (BIC), section 4.4.1.

1.4 The curse of dimensionality

- curse: malédiction, fléau ...
- polynomial fitting: replace x by a vector $\boldsymbol{x}$ of dimension D. The number of unknowns becomes D^M.
- *Not all the intuitions developed in spaces of low dimensionality will generalize to spaces of many dimensions*

Section 1.5: Decision theory

Decision theory - introduction

- The decision problem:
  - given x, predict t according to a probabilistic model $p(x,t)$
- For now: binary classification: $t \in \{0,1\} \Leftrightarrow \{C_1, C_2\}$
- Important quantity: $p(C_k|x)$

$$p(C_k|x) = \frac{p(x,C_k)}{p(x)} = \frac{p(x,C_k)}{\sum_{k=1}^{2} p(x,C_k)}$$

$\Rightarrow$ getting $p(x,C_i)$ is the (central!) inference problem

$$= \frac{p(x|C_k)p(C_k)}{p(x)}$$

$\propto$ likelihood $\times$ prior

- Intuition: choose k that maximizes $p(C_k|x)$?

Decision theory - binary classification

- Decision region: $\mathcal{R}_i = \{x : \text{pred}(x) = C_i\}$
- Probability of misclassification:

$$
\begin{aligned}
p(\text{mis}) &= p(x \in \mathcal{R}_1, C_2) + p(x \in \mathcal{R}_2, C_1) \\
&= \int_{\mathcal{R}_1} p(x, C_2) dx + \int_{\mathcal{R}_2} p(x, C_1) dx
\end{aligned}
$$

$\Rightarrow$ In order to minimize, affect x to $\mathcal{R}_1$ if:

$$
\begin{aligned}
& p(x, C_1) > p(x, C_2) \\
& \Leftrightarrow p(C_1|x)p(x) > p(C_2|x)p(x) \\
& \Leftrightarrow p(C_1|x) > p(C_2|x)
\end{aligned}
$$

- Similarly, for k classes: minimize $\sum_j \int_{\mathcal{R}_j} \Big(\sum_{k \neq j} p(x, C_k) \Big) dx$

$\Rightarrow \text{pred}(x) = \text{argmax}_k p(C_k|x)$

Decision theory - loss-sensitive decision

- Cost/Loss of a decision: L_{kj} = predict C_j while truth is C_k.
- Loss-sensitive decision $\Rightarrow$ minimize the expected loss:

$$E[L] = \sum_j \int_{\mathcal{R}_j} \Big(\sum_k L_{kj} p(x, C_k) \Big) dx$$

- Solution: for each x, choose the class C_j that minimizes:

$$\sum_k L_{kj} p(x, C_k) \propto \sum_k L_{kj} p(C_k|x)$$

$\Rightarrow$ straightforward when we know $p(C_k|x)$

Decision theory - loss-sensitive decision

- Typical example = medical diagnosis:
  - $C_k = \{1, 2\} \Leftrightarrow \{\text{sick}, \text{healthy}\}$
  - $L = \begin{bmatrix} 0 & 100 \\ 1 & 0 \end{bmatrix} \Rightarrow$ strong cost of "missing" a diseased person
- Expected loss:

$$
\begin{aligned}
E[L] &= \int_{\mathcal{R}_2} L_{1,2} p(x, C_1) dx + \int_{\mathcal{R}_1} L_{2,1} p(x, C_2) dx \\
&= \int_{\mathcal{R}_2} 100 \times p(x, C_1) dx + \int_{\mathcal{R}_1} p(x, C_2) dx
\end{aligned}
$$

- Note: minimizing the probability of misclassification:

$$
p(\text{mis}) = \int_{\mathcal{R}_1} p(x, C_2) dx + \int_{\mathcal{R}_2} p(x, C_1) dx
$$

corresponds to minimizing the 0/1 loss: $L = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Decision theory - the "reject option"

- For the 0/1 loss[1], $\text{pred}(x) = \text{argmax}_k p(C_k|x)$
  - Note: K classes $\Rightarrow 1/K \leq \max_k p(C_k|x) \leq 1$
- When $\max_k p(C_k|x) \to 1/K$ the confidence in the prediction decreases.
  - classes tend to become as likely
- "Reject option": make a decision provided $\max_k p(C_k|x) > \sigma$
  $\Rightarrow$ the value of σ controls the amount of rejection:
  - $\sigma = 1$: systematic rejection
  - $\sigma < 1/K$: no rejection
- Motivation: switch between automatic/human decision
- Illustration in Figure 1.26 page 42

[1] For a general loss matrix, see exercice 1.24

Decision theory - regression setting

- The regression setting: quantitative target $t \in \mathbb{R}$
- Typical regression loss-function: $L(t, y(x)) = (y(x) - t)^2$
  - the squared loss
- The decision problem = minimize the expected loss:

$$E[L] = \int_{\mathcal{X}} \int_{\mathbb{R}} (y(x) - t)^2 p(x, t) dx dt$$

- Solution: $y(x) = \int_{\mathbb{R}} t p(t|x) dt$
  - this is known as the regression function
  - intuitively appealing: conditional average of t given x
    - illustration in figure 1.28, page 47

- Note: general class of loss functions $L(t, y(x)) = |y(x) - t|^q$
  - $q = 2$ is analytically convenient (derivable and continuous)

Decision theory - regression setting

- Derivation:

$$
\begin{aligned}
E[L] &= \int_{\mathcal{X}} \int_{\mathbb{R}} (y(x) - t)^2 p(x,t) dx dt \\
&= \int_{\mathcal{X}} \Big[\int_{\mathbb{R}} (y(x) - t)^2 p(t|x) dt \Big] p(x) dx
\end{aligned}
$$

$\Rightarrow$ for each x, find $y(x)$ that minimizes $\int_{\mathbb{R}} (y(x) - t)^2 p(t|x) dt$

- Derivating with respect to $y(x)$ gives: $2 \int_{\mathbb{R}} (y(x) - t) p(t|x) dt$
- Setting to zero leads to:

$$
\begin{aligned}
\int_{\mathbb{R}} y(x) p(t|x) dt &= \int_{\mathbb{R}} t p(t|x) dt \\
y(x) &= \int_{\mathbb{R}} t p(t|x) dt
\end{aligned}
$$

Decision theory - inference and decision

2 (or 3) different approaches to the decision problem:

1. rely on a probabilistic model, with 2 flavours:
   1.1 generative:
      - use a generative model to infer $p(x|C_k)$
      - combine with priors $p(C_k)$ to get $p(x, C_k)$ and eventually $p(C_k|x)$

   1.2 discriminative: infer directly $p(C_k|x)$
      - this is sufficient for the decision problem
2. learn a discriminant function $f(x)$
   - directly map input to class labels
   - for binary classification, $f(x)$ is typically defined as the sign $(+1/\text{-}1)$ of an auxiliary function

(Note: similar discussion for regression)

Decision theory - inference and decision

Pros and Cons:

- probabilistic generative models:
  - pros: acess to $p(x) \rightarrow$ easy detection of outliers
    - i.e., low-confidence predictions
  - cons: estimating the joint probability $p(x, C_k)$ can be computational and data demanding
- probabilistic discrimative models:
  - pros: less demanding than the generative approach
    - see figure 1.27, page 44
- discriminant functions:
  - pros: a single learning problem (vs inference + decision)
  - cons: no access to $p(C_k|x)$
    - ... which can have many advantages in practice for (e.g.) rejection and model combination – see page 45

Section 1.6: Information theory

Information theory - Entropy

- Consider a discrete random variable X
- We want to define a measure $h(x)$ of surprise/information of observing $X = x$
- Natural requirements:
  - if $p(x)$ is low (resp. high), $h(x)$ should be high (resp.low)
    - $h(x)$ should be a monotonically decreasing function of $p(x)$
  - if X and Y are unrelated, $h(x, y)$ should be $h(x) + h(y)$
    - i.e., if X and Y are independent, that is $p(x, y) = p(x)p(y)$

  $\Rightarrow$ this leads to $h(x) = -\log p(x)$
- Entropy of the variable X:

$$H[X] = E[h(X)] = -\sum_x p(x) \log(p(x))$$

(Convention: $p \log p = 0$ if $p = 0$)

Information theory - Entropy

Some remarks:

- $H[X] \geq 0$ since $p \in [0,1]$ (hence $p \log p \leq 0$)
- $H[X] = 0$ if $\exists\, x$ s.t. $p(x) = 1$
- Maximum entropy distribution $=$ uniform distribution
  - optimization problem: maximize $H[X] + \lambda\big(\sum_{x_i} p(x_i) - 1\big)$
  - derivating w.r.t. $p(x_i)$ shows they must be constant
  - hence $p(x_i) = 1/M, \forall x_i \Rightarrow H[X] = \log(M)$

$\Rightarrow$ we therefore have $0 \leq H[X] \leq \log(M)$

- $H[X] =$ lower bound on the # of bits required to (binarily) encode the values of X (using $\log_2$ in the defintion of H)
  - trivial code of length $\log_2(M)$ (ex: $M = 8$, messages of size 3)
    - no "clever" coding scheme for uniform distributions
  - for non-uniform distributions, optimal coding schemes can be designed
    - high probability values $\Rightarrow$ short codes
  - see illustration in page 50

Information theory - Entropy

For continuous random variables:

- differential entropy: $H[X] = -\int p(x) \ln p(x) dx$
  - because $p(x)$ can be > 1, care must be taken when transposing properties of the discrete entropy
  - in particular, can be negative (if $X \hookrightarrow \mathcal{U}(0, 1/2)$: $H[X] = -\ln 2$)
- Given (μ, σ), maximum entropy distribution $p(x) = \mathcal{N}(\mu, \sigma^2)$
  - optimization problem: maximize $H[X]$ with μ, σ equality constraint + normalization constraint
  - entropy: $H[X] = 1/2\big(1 + \ln(2\pi\sigma^2)\big)$
- Conditional entropy of y given x:

$$H[Y|X] = -\int\int p(x,y) \ln p(y|x) dx dy$$

$\Rightarrow$ we have easily $H[X,Y] = H[Y|X] + H[X]$

(natural interpretation with the notion of information)

Information theory - KL divergence

- Kullback-Leibler divergence between distributions p and q:

$$KL(p||q) = -\int p(x) \ln q(x) dx - \left(-\int p(x) \ln p(x) dx \right)$$
$$= -\int p(x) \ln \frac{q(x)}{p(x)} dx$$

- $KL(p||q) \neq KL(q||p)$
- $KL(p||p) = 0$
- $KL(p||q) \geq 0$ (next slide)

$\Rightarrow$ measures the difference between the "true" distribution p and the distribution q

(Information therory interpretation: amount of additional information required to encode the values of X using $q(x)$ instead of $p(x)$)

Information theory - KL divergence

- A function is convex iff every cord lies above the function
  - illustration in figure 1.31, page 56
- Jensen's inequality for convex functions:

$$E[f(x)] \geq f\big(E[x]\big)$$

  (strict inequality for strictly convex functions)

- When applied to $KL(p||q)$:

$-\ln x$ 볼록함수
젠슨 부등식에 의해

$$\begin{aligned} KL(p||q) &= -\int p(x)\ln\frac{q(x)}{p(x)}dx \\ &> -\ln\int p(x)\times\frac{q(x)}{p(x)}dx \quad \text{(because } -\ln \text{ is stricly convex)} \\ &= -\ln\underline{\int q(x)dx} = -\ln 1 = 0 \end{aligned}$$

  Moreover, straightforward to see that $KL(p||p) = 0$

- Conclusion: $KL(p||q) \geq 0$, with equality if $p = q$

Information theory - KL divergence: illustration

- Data generated by an (unknown) distribution $p(x)$
- We want to fit a parametric probabilistic model $q(x|\theta) = q_\theta(x)$
  $\Rightarrow$ i.e., we want to minimize $KL(p||q_\theta)$
- Data available: observations $(x_1, \ldots, x_N)$:

$-\ln x$ 은 볼록함수

$$
\begin{aligned}
KL(p||q_\theta) &= -\ln \int p(x) \times \ln \frac{q(x|\theta)}{p(x)} dx \\
&\simeq -\sum_{i=1}^{N} \ln \frac{q(x_i|\theta)}{p(x_i)} \\
&= \sum_{i=1}^{N} \Big(-\ln q(x_i|\theta) + \ln p(x_i) \Big)
\end{aligned}
$$

$\Rightarrow$ it follows that minimizing $KL(p||q_\theta)$ corresponds to maximizing $\sum_{i=1}^{N} \ln q(x_i|\theta) =$ log-likelihood

Information theory - Mutual information

- Mutual information: $I[X,Y] = KL(p(X,Y)||p(X),p(Y))$
- Quantifies the amount of independence between X and Y
  - $I[X,Y] = 0 \Leftrightarrow p(X,Y) = p(X)p(Y)$
- We have:

$$
\begin{aligned}
I[x,y] &= -\int\int p(x,y) \ln \frac{p(x)p(y)}{p(x,y)} dxdy \\
&= -\int\int p(x,y) \ln \frac{p(x)p(y)}{p(x|y)p(y)} dxdy \\
&= -\int\int p(x,y) \ln \frac{p(x)}{p(x|y)} dxdy \\
&= -\int\int p(x,y) \ln p(x) dxdy - \Big(-\int\int p(x,y) \ln p(x|y) dxdy \Big) \\
&= H[X] - H[X|Y]
\end{aligned}
$$

- Conclusion: $I[X,Y] = H[X] - H[X|Y] = H[Y] - H[Y|X]$
  - $I[X,Y] =$ reduction of the uncertainty about X obtained by telling the value of Y (that is, 0 for independent variables)