Machine Learning

Bayesian Learning

Lecturer: Duc Dung Nguyen, PhD.
Contact: nddung@hcmut.edu.vn

Faculty of Computer Science and Engineering
Hochiminh city University of Technology

Contents

Linear Prediction

Linear supervised learning

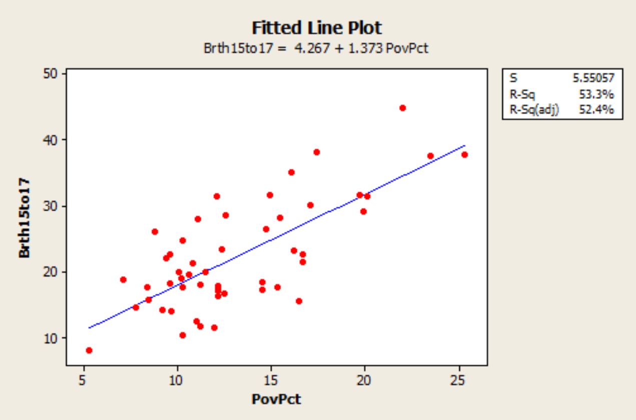
- Many real processes can be **approximated** with linear models
- Linear regression often appears as a **module** of larger systems
- Linear problems can be solved **analytically**
- Linear prediction provides an introduction to many of the **core concepts** in machine learning.

Linear Prediction

Energy demand prediction

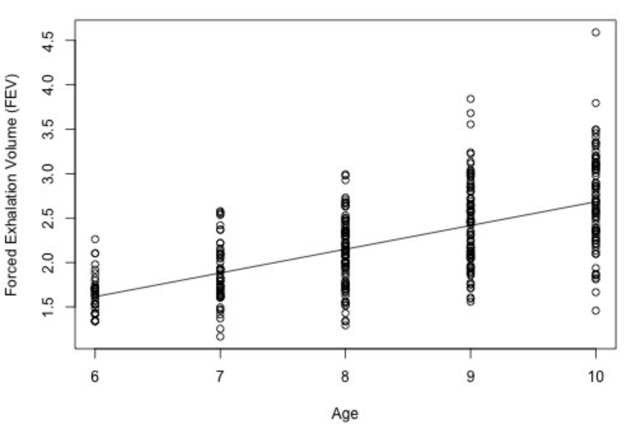
Wind speed	People inside building	Energy requirement
100	2	5
50	42	25
45	31	22
60	35	18

Teen Birth Rate and Poverty Level Data

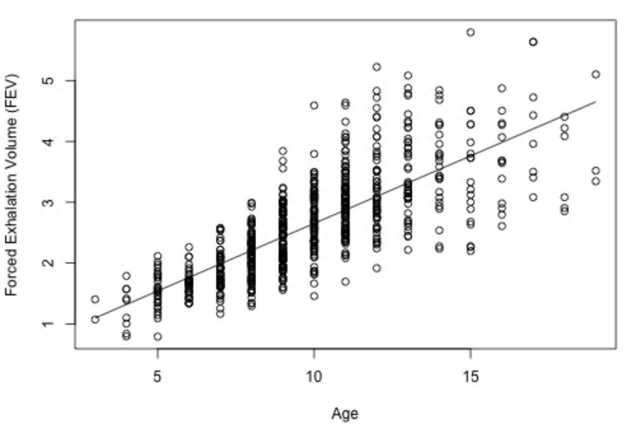


Linear Prediction

Lung Function in 6 to 10 Year Old Children



Lung Function in 6 to 10 Year Old Children



Linear Prediction

- In general the linear model is expressed as follows

$$\hat{y}_i = \sum_{j=1}^{d} x_{ij}\theta_j$$

- In matrix form

$$\hat{\mathbf{y}} = \mathbf{X}\theta$$

Linear Prediction

- We can use optimization approach

$$\mathbf{J}(\theta) = (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}})$$

- Least squares estimates
- Probabilistic approach

Bayesian Learning

Bayesian Learning

- It involves **direct manipulation of probabilities** in order to find correct hypotheses.
- The quantities of interest are governed by **probability distributions**.
- Optimal decisions can be made by **reasoning about those probabilities**.

Bayesian Learning

- Bayesian learning algorithms are among the most **practical approaches** to certain type of learning problems
- Provide a useful perspective for **understanding many learning algorithms** that do not explicitly manipulate probabilities.

- Each training example can **incrementally** decrease or increase the estimated probability that a hypothesis is correct.
- **Prior knowledge** can be combined with observed data to determine the final probability of a hypothesis
- **Hypotheses with probabilities** can be accommodated
- New instances can be classified by **combining multiple hypotheses** weighted by the probabilities.

Bayes Theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \tag{1}$$

- $P(h)$: prior probability of hypothesis **h**
- $P(D)$: prior probability of training data **D**
- $P(h|D)$: probability that **h** holds given **D**
- $P(D|h)$: probability that **D** is observed given **h**

- **Maximum A-posteriori hypothesis (MAP)**:

$$h_{MAP} = \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} P(D|h)P(h) \tag{2}$$

$P(h)$ is **not a uniform distribution** over H.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \tag{3}$$

- **Maximum Likelihood hypothesis (ML)**:

$$h_{ML} = \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} P(D|h) \quad (4)$$

If $P(h)$ is **a uniform distribution** over H.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad (5)$$

Bayes Theorem

- **0.008** of the population have cancer

Bayes Theorem

- **0.008** of the population have cancer
- Only **98%** patients are correctly classified as **positive**

Bayes Theorem

- **0.008** of the population have cancer
- Only **98%** patients are correctly classified as **positive**
- Only **97%** non-patiants are correctly classified as **negative**

Bayes Theorem

- **0.008** of the population have cancer
- Only **98%** patients are correctly classified as **positive**
- Only **97%** non-patiants are correctly classified as **negative**
- *Would a person with a positive result have cancer or not?*

$$P(cancer|\oplus) >< P(\neg cancer|\oplus)$$

- **Maximum A-posteriori hypothesis (MAP)**:

$$
\begin{aligned}
h_{MAP} &= \underset{h \in (cancer, \neg cancer)}{\arg\max} P(h|\oplus) \\
&= \underset{h \in (cancer, \neg cancer)}{\arg\max} P(\oplus|h)P(h)
\end{aligned} \tag{6}
$$

Bayes Theorem

- $P(cancer) = .008 \rightarrow P(\neg cancer) = .992$

Bayes Theorem

- $P(cancer) = .008 \rightarrow P(\neg cancer) = .992$
- $P(\oplus | cancer) = .98$

- $P(cancer) = .008 \rightarrow P(\neg cancer) = .992$
- $P(\oplus | cancer) = .98$
- $P(\ominus | \neg cancer) = .97 \rightarrow P(\oplus | \neg cancer) = .03$

Bayes Theorem

- $P(cancer) = .008 \rightarrow P(\neg cancer) = .992$
- $P(\oplus|cancer) = .98$
- $P(\ominus|\neg cancer) = .97 \rightarrow P(\oplus|\neg cancer) = .03$
- $P(cancer|\oplus) \approx P(\oplus|cancer)p(cancer) = .0078$

Bayes Theorem

- $P(cancer) = .008 \rightarrow P(\neg cancer) = .992$
- $P(\oplus|cancer) = .98$
- $P(\ominus|\neg cancer) = .97 \rightarrow P(\oplus|\neg cancer) = .03$
- $P(cancer|\oplus) \approx P(\oplus|cancer)p(cancer) = .0078$
- $P(\neg cancer|\oplus) \approx P(\oplus|\neg cancer)P(\neg cancer) = .0298$

- **Maximum A-posteriori hypothesis (MAP)**:

$$\begin{aligned} h_{MAP} &= \underset{h \in (cancer, \neg cancer)}{\arg\max} P(h|\oplus) \\ &= \underset{h \in (cancer, \neg cancer)}{\arg\max} P(\oplus|h)P(h) \\ &= \neg cancer \end{aligned} \tag{7}$$

Bayes Optimal Classifier

- What is the most **probable hypothesis** given the training data?
- What is the most **probable classification** of a new instance given the training data?

- Hypothesis space $= \{h_1, h_2, h_3\}$
- Posterior probabilities $= \{.4, .3, .3\}$ (h_1 is h_{MAP})
- New instance x is classified positive by h_1 and negative by h_2 and h_3

What is the most probable classification of x?

Bayes Optimal Classifier

- The most probable classification of a new instance is obtained by combining the predictions of **all hypotheses** *weighted by their posterior probabilities*:

$$\arg\max_{c \in C} P(c|D) = \arg\max_{c \in C} \sum_{h \in H} P(c|h).P(h|D) \tag{8}$$

Example	**Sky**	**AirTemp**	**Humidity**	**Wind**	**Water**	**Forecast**	**EnjoySport**
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes
5	Cloudy	Warm	High	Weak	Cool	Same	Yes
6	Sunny	Cold	High	Weak	Cool	Same	No
7	Sunny	Warm	Normal	Strong	Warm	Same	?
8	Sunny	Warm	Low	Strong	Cool	Same	?

- Each instance **x** is described by a conjunction of attribute values $< a_1, a_2, ..., a_n >$
- The target function $f(x)$ can take on any value from a finite set C
- It is to assign the most probable target value to a new instance

Naive Bayes Classifier

$$C_{MAP} = \arg\max_{c \in C} P(c|a_1, a_2, ..., a_n) \\ = \arg\max_{c \in C} P(a_1, a_2, ..., a_n|c)P(c) \tag{9}$$

Naive Bayes Classifier

$$\begin{aligned} C_{MAP} &= \arg\max_{c \in C} P(c|a_1, a_2, ..., a_n) \\ &= \arg\max_{c \in C} P(a_1, a_2, ..., a_n|c)P(c) \end{aligned} \tag{10}$$

$$C_{NB} = \arg\max_{c \in C} \prod_{i=1,n} P(a_i|c)P(c)$$

assuming that $a_1, a_2, ..., a_n$ are independent given c

Naive Bayes Classifier

Example	**Sky**	**AirTemp**	**Humidity**	**Wind**	**Water**	**Forecast**	*EnjoySport*
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes
5	Cloudy	Warm	High	Weak	Cool	Same	Yes
6	Sunny	Cold	High	Weak	Cool	Same	No
7	Sunny	Warm	Normal	Strong	Warm	Same	?
8	Sunny	Warm	Low	Strong	Cool	Same	?

Estimating probabilities:

- Probability: the fraction of times the event is observed to occur over the total number of opportunities n_c/n
- **What if the fraction is too small, or even zero?**

Estimating probabilities:

$$\frac{n_c + mp}{n + m} \tag{11}$$

- n: total number of training examples of a particular class.
- n_c: number of training examples having a particular attribute value in that class.
- m: equivalent sample size
- p: prior estimate of the probability (equals $1/k$ where k is the number of possible values of the attribute)

Learning to classify text:

$$C_{NB} = \arg\max_{c \in C} \prod_{i=1,n} P(a_i = w_k | c).P(c)$$

Learning to classify text:

$$C_{NB} = \arg\max_{c \in C} \prod_{i=1,n} P(a_i = w_k|c).P(c) = \arg\max_{c \in C} \prod_{i=1,n} P(w_k|c).P(c) \tag{12}$$

assuming that all words have equal chance occurring in every position