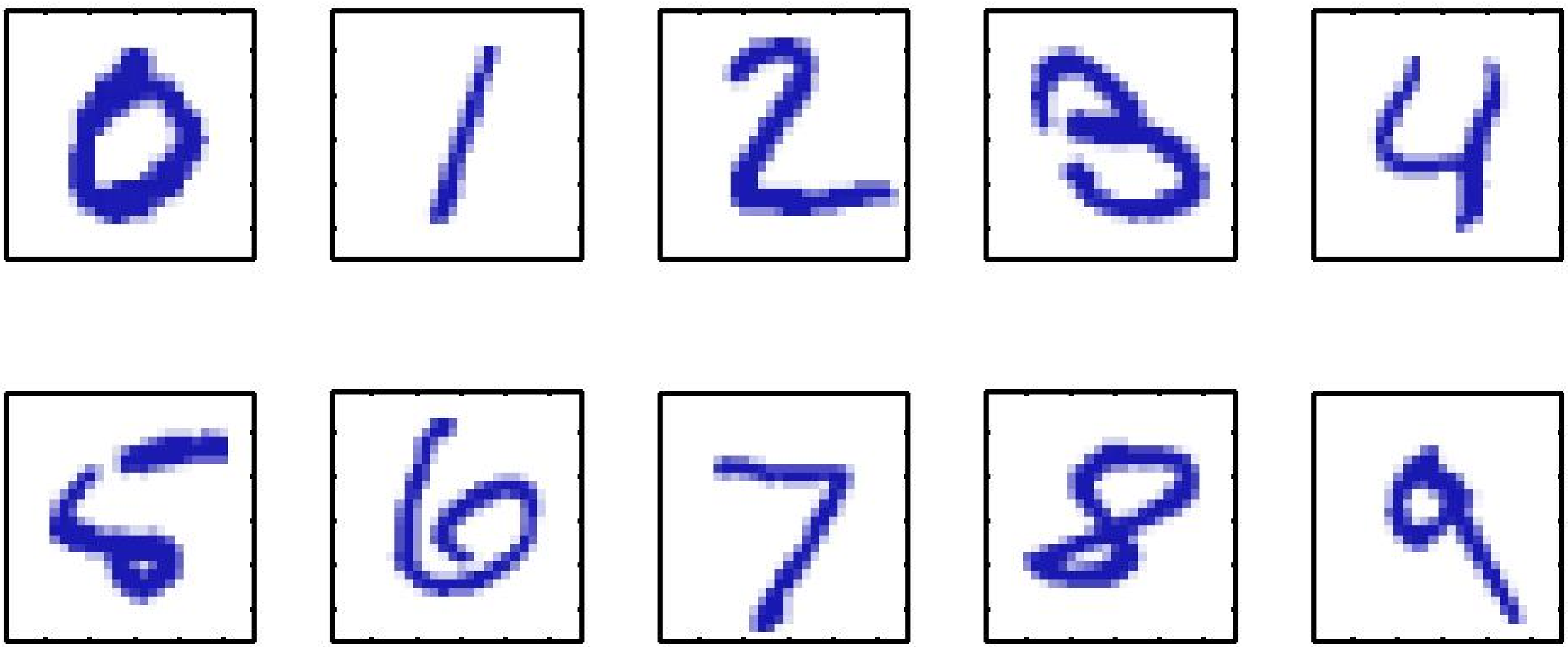
CSE6363 Machine Learning, Prof. Chris Ding

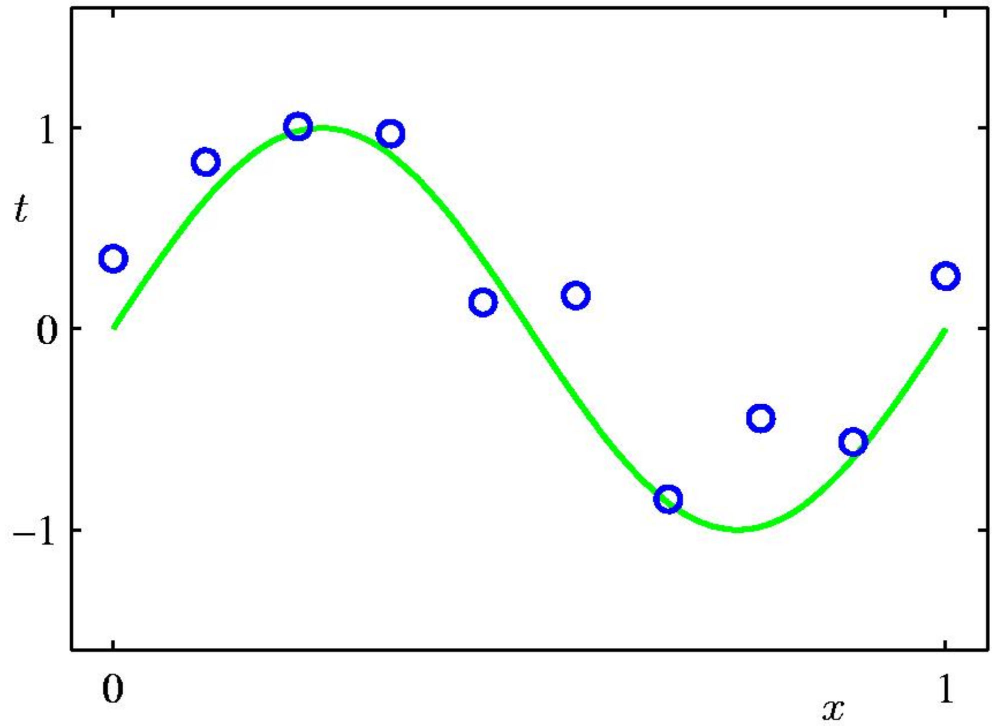
Fitting Data Using Polynomial
Linear Regression
Regularization
Probability
Bernoulli Distribution, Binomial Distribution, Multinomial Distribution
Normal Distribution
Linear Regression as Maximum Likelihood Estimation from Normal Distribution
Bayes Theorem
Naïve Bayes Classification

Example

Handwritten Digit Recognition

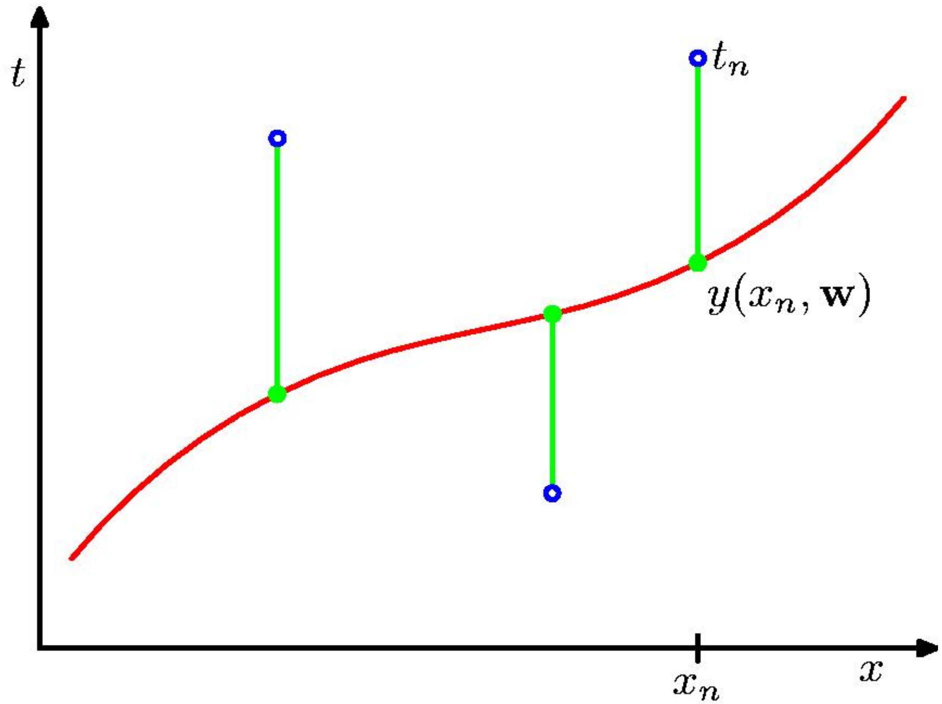


Polynomial Curve Fitting



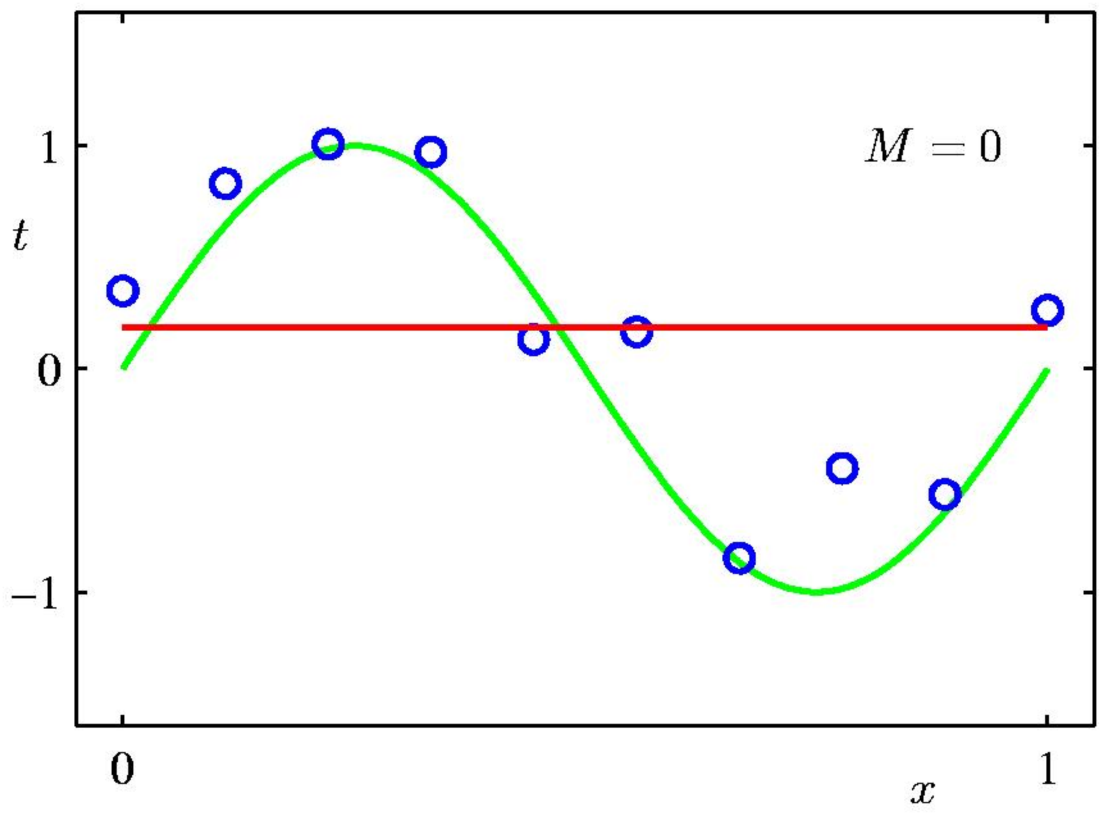
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Sum-of-Squares Error Function

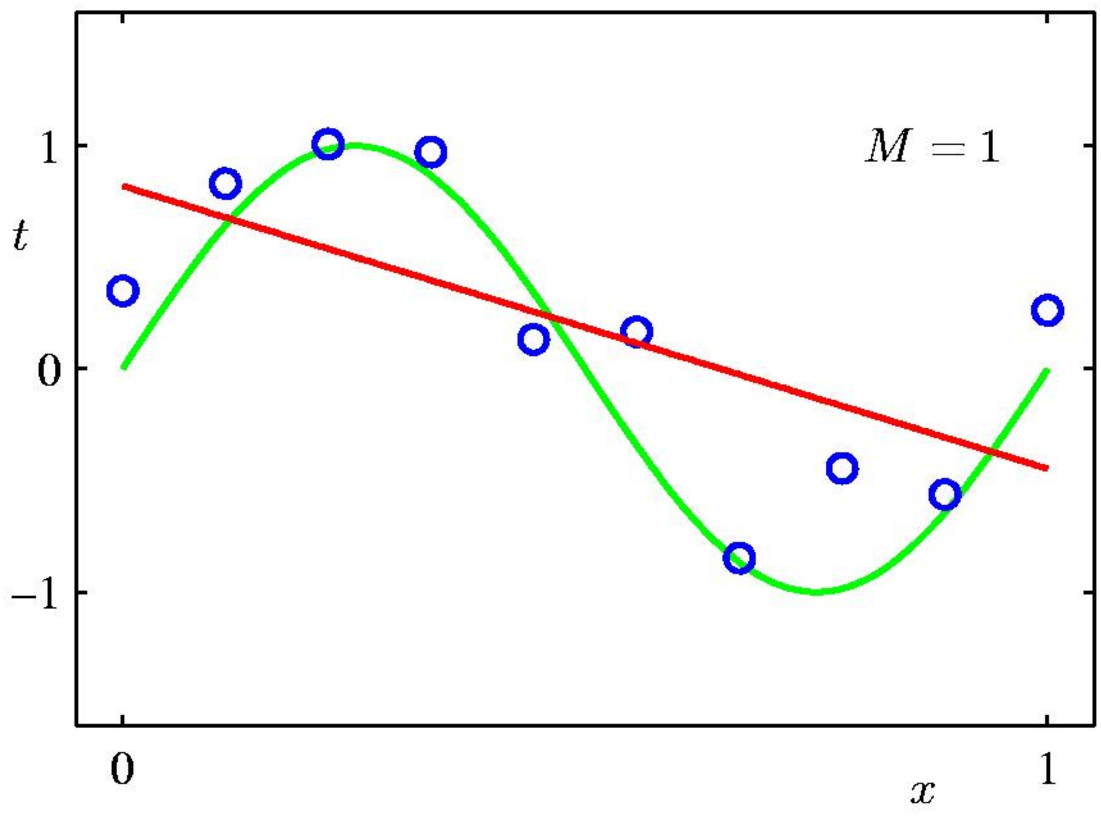


$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

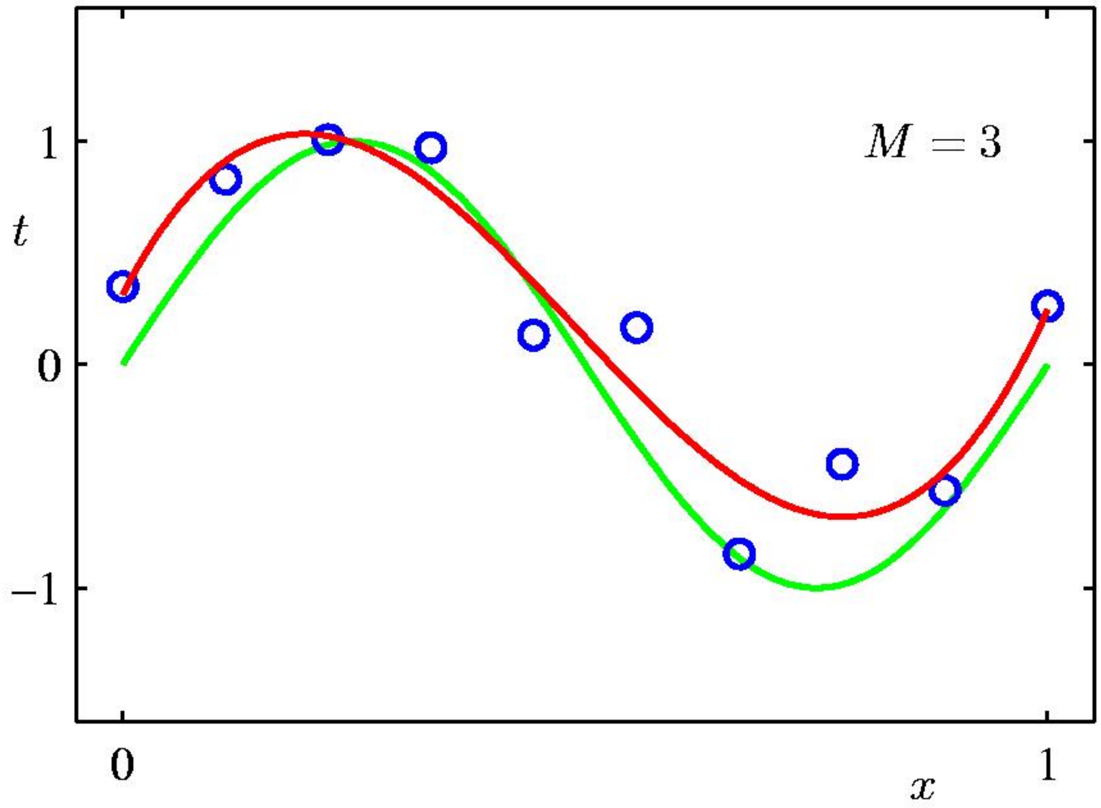
0^{th} Order Polynomial



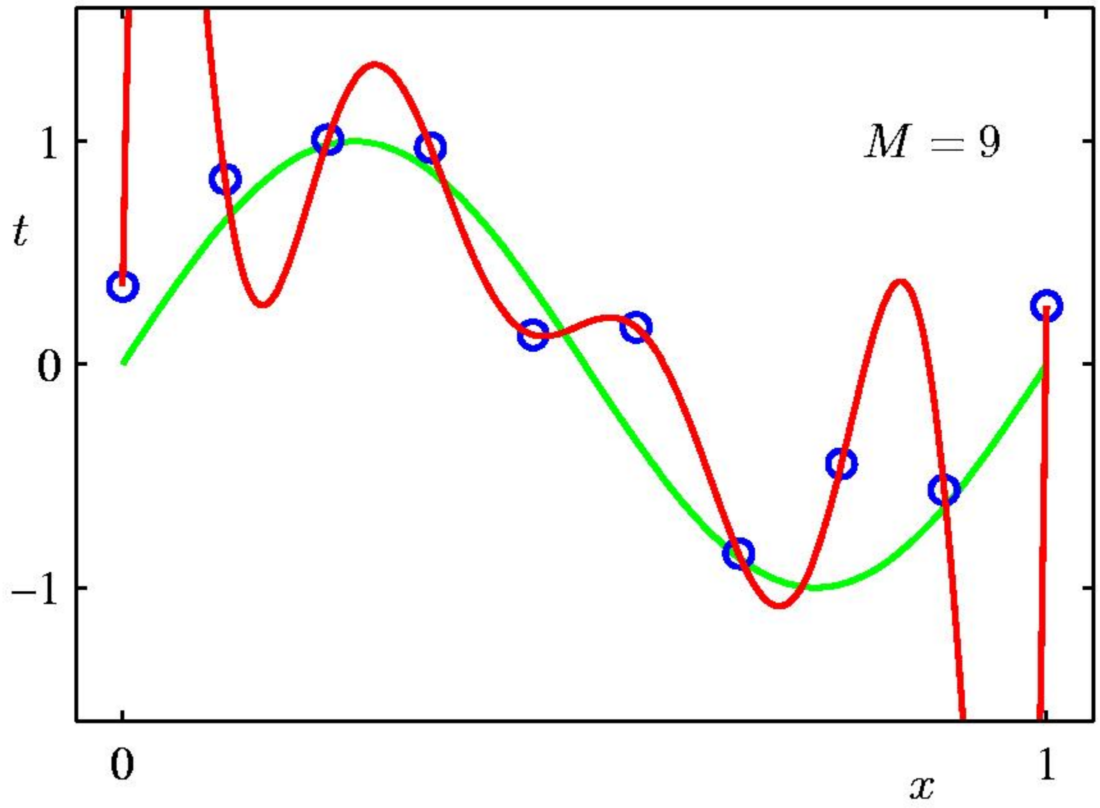
1^{st} Order Polynomial



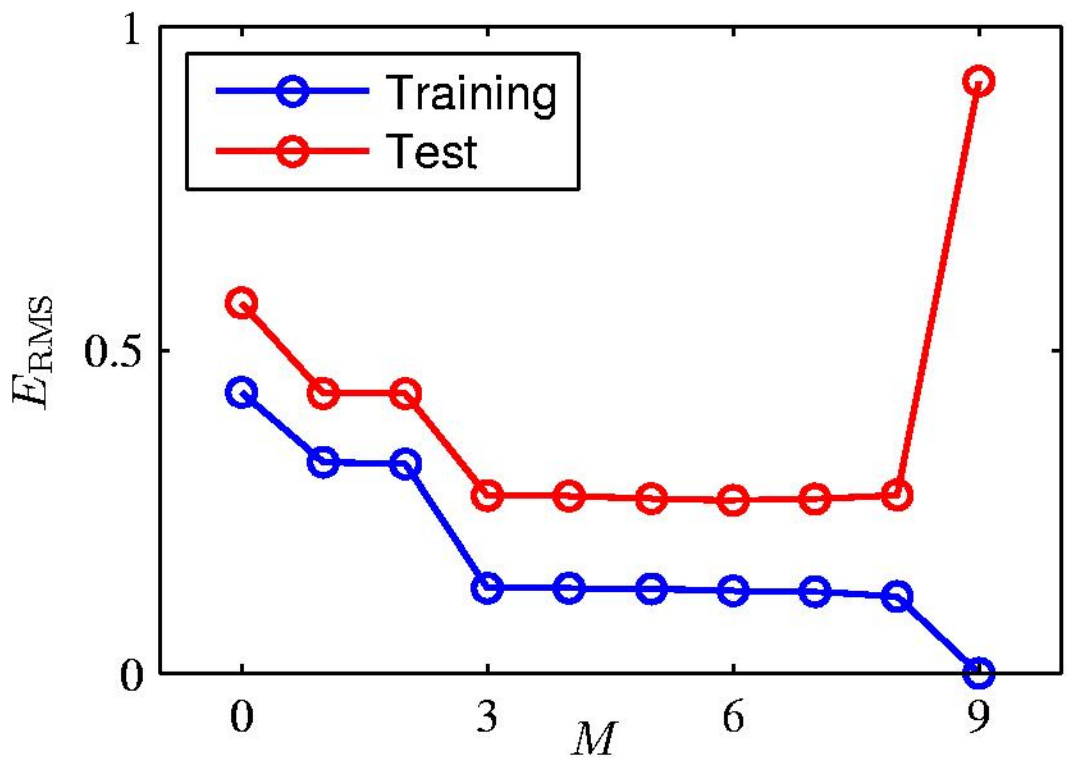
3^{rd} Order Polynomial



9th Order Polynomial



Over-fitting



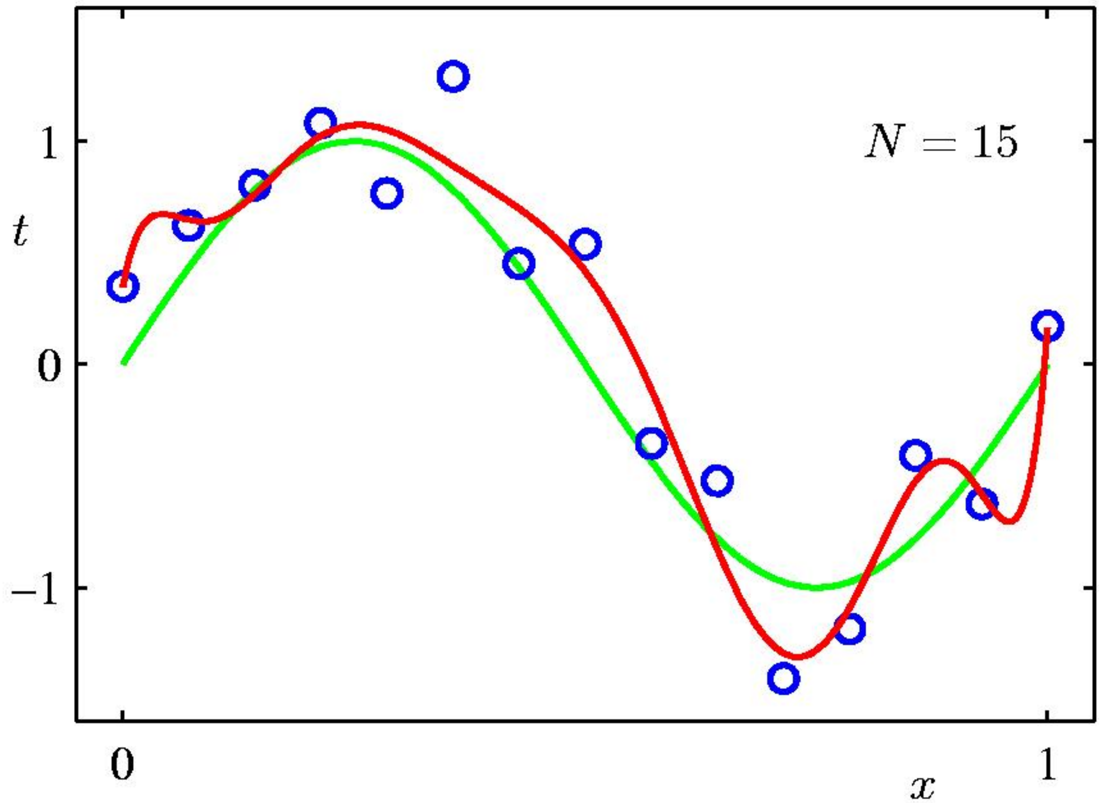
Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^{\star})/N}$

Polynomial Coefficients

	$M = 0$	$M = 1$	$M = 3$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

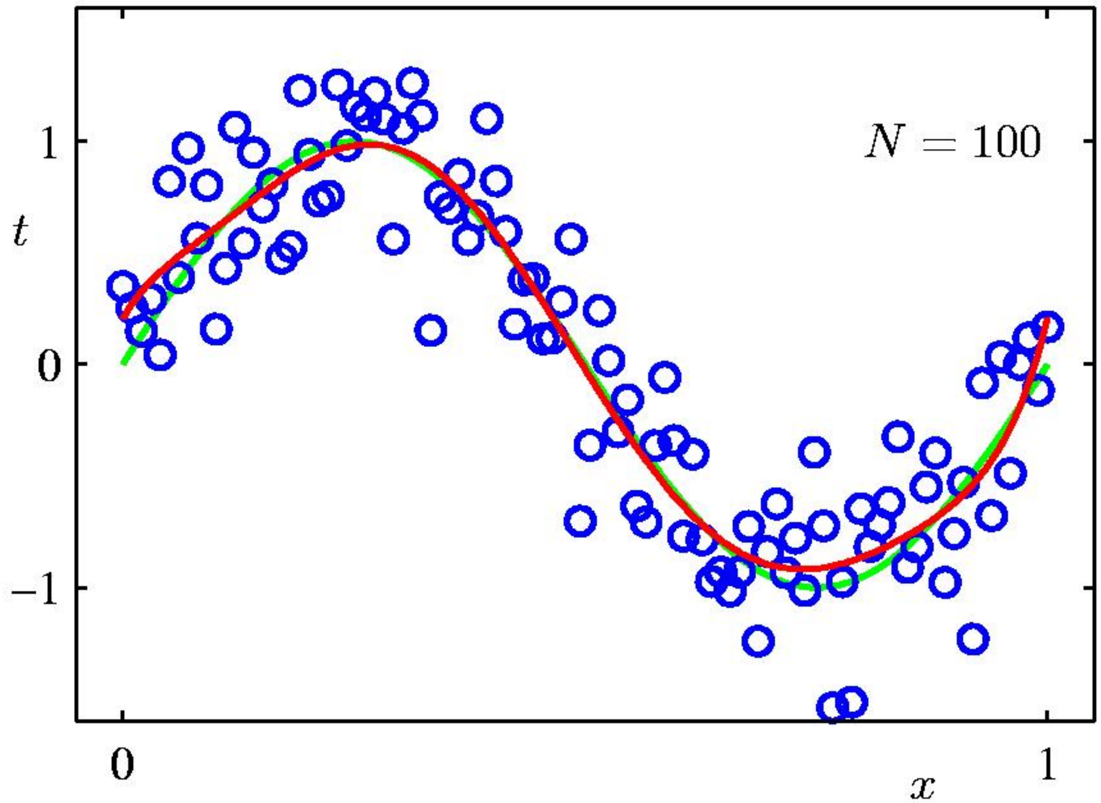
Data Set Size: $N = 15$

9^{th} Order Polynomial



Data Set Size: $N = 100$

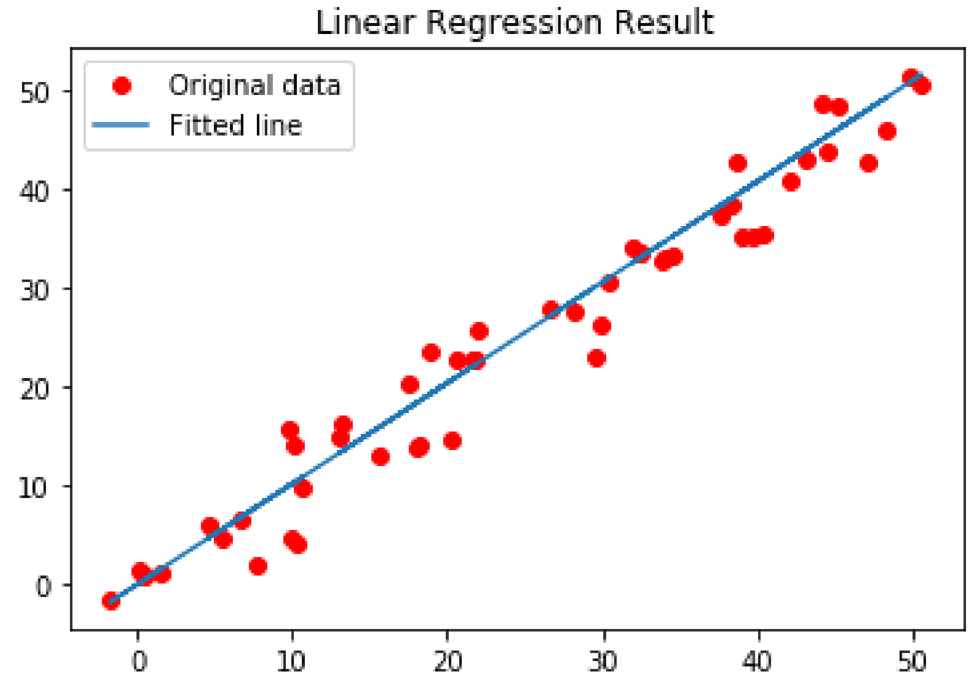
9th Order Polynomial



Regularization

Penalize large coefficient values

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



Linear Regression Result
Original data
Fitted line
50
40
30
20
10
0
0
10
20
30
40
50

Heat Flux	Skin Temperature
10.858	31.002
10.617	31.021
10.183	31.058
9.7003	31.095
9.652	31.133
10,086	31,188
9.459	31,226
8.3972	31,263
7.6251	31,319
7.1907	31,356
7.046	31.412
6.9494	31.468
6.7081	31.524

Heat Flux	Skin Temperature
6.3221	31.581
6.0325	31.618
5.7429	31.674
5.5016	31.712
5.2603	31.768
5,1638	31,825
5,0673	31.862
4.9708	31.919
4,8743	31.975
4.7777	32.013
4.7295	32.07
4.633	32.126
4.4882	32.164

Heat Flux	Skin Temperature
4.3917	32,221
4.2951	32.259
4.2469	32.296
4.0056	32.334
3.716	32.391
3,523	32.448
3,4265	32.505
3,3782	32,543
3,4265	32.6
3.3782	32.657
3.3299	32,696
3,3299	32.753
3.4265	32.791

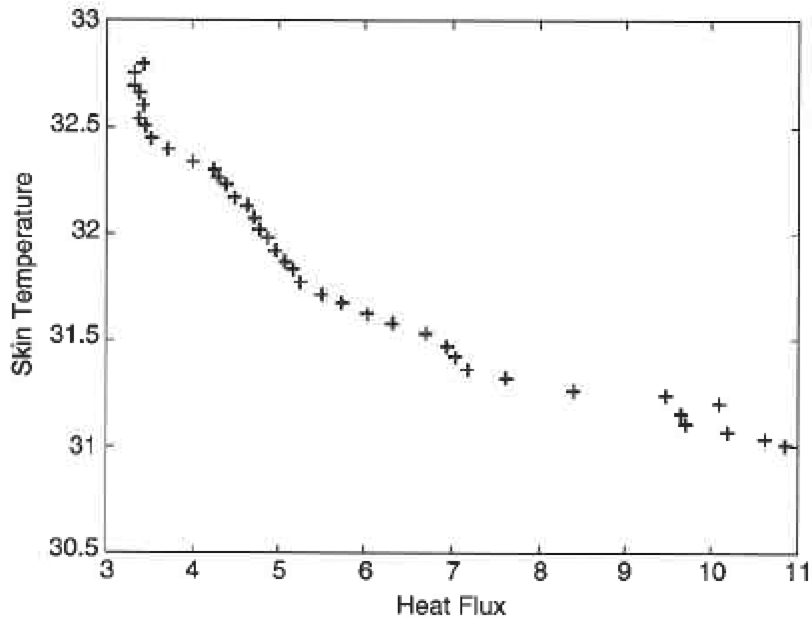


Figure D.1. Measurements of heat flux and skin temperature of a person.

D.2.1 Least Square Method

Suppose we wish to fit the following linear model to the observed data:

$$f(x) = \omega_1 x + \omega_0, \tag{D.3}$$

where ω_0 and ω_1 are parameters of the model and are called the **regression coefficients**. A standard approach for doing this is to apply the **method of least squares**, which attempts to find the parameters (ω_0, ω_1) that minimize the sum of the squared error

$$SSE = \sum_{i=1}^{N}[y_i - f(x_i)]^2 = \sum_{i=1}^{N}[y_i - \omega_1 x - \omega_0]^2, \tag{D.4}$$

which is also known as the **residual sum of squares**.

This optimization problem can be solved by taking the partial derivative of E with respect to ω_0 and ω_1, setting them to zero, and solving the corresponding system of linear equations.

$$\begin{aligned}
\frac{\partial E}{\partial \omega_0} &= -2\sum_{i=1}^{N}[y_i - \omega_1 x_i - \omega_0] = 0 \\
\frac{\partial E}{\partial \omega_1} &= -2\sum_{i=1}^{N}[y_i - \omega_1 x_i - \omega_0]x_i = 0
\end{aligned} \tag{D.5}$$

These equations can be summarized by the following matrix equation, which is also known as the **normal equation**:

$$\begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \begin{pmatrix} \omega_0 \\ \omega_1 \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}. \tag{D.6}$$

Since $\sum_i x_i = 229.9$, $\sum_i x_i^2 = 1569.2$, $\sum_i y_i = 1242.9$, and $\sum_i x_i y_i = 7279.7$, the normal equations can be solved to obtain the following estimates for the parameters.

$$
\begin{aligned}
\begin{pmatrix} \hat{\omega}_0 \\ \hat{\omega}_1 \end{pmatrix} &= \begin{pmatrix} 39 & 229.9 \\ 229.9 & 1569.2 \end{pmatrix}^{-1} \begin{pmatrix} 1242.9 \\ 7279.7 \end{pmatrix} \\
&= \begin{pmatrix} 0.1881 & -0.0276 \\ -0.0276 & 0.0047 \end{pmatrix} \begin{pmatrix} 1242.9 \\ 7279.7 \end{pmatrix} \\
&= \begin{pmatrix} 33.1699 \\ -0.2208 \end{pmatrix}
\end{aligned}
$$

$$f(x) = 33.17 - 0.22x.$$

Figure D.2 shows the line corresponding to this model.

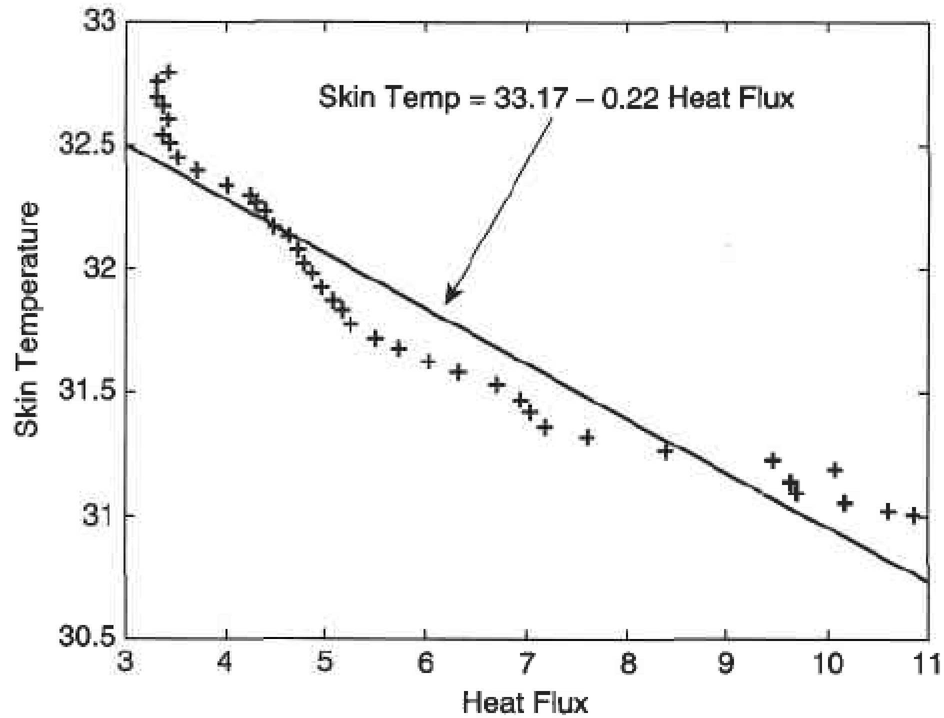
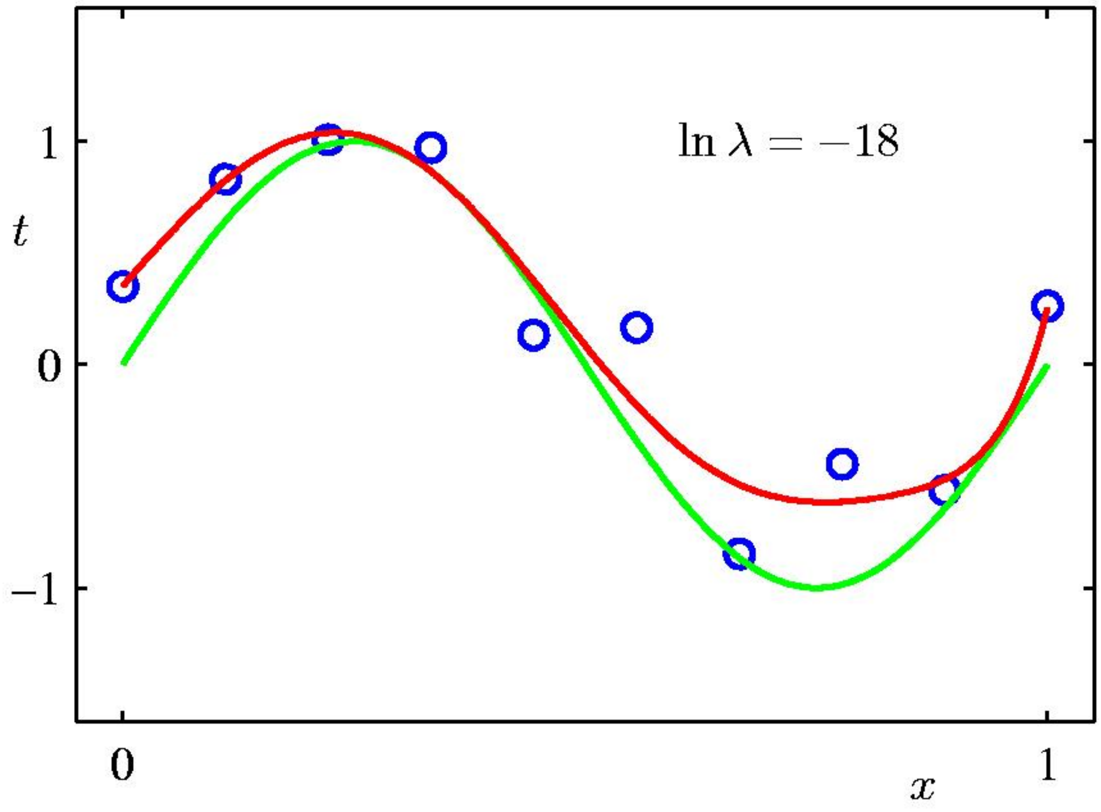
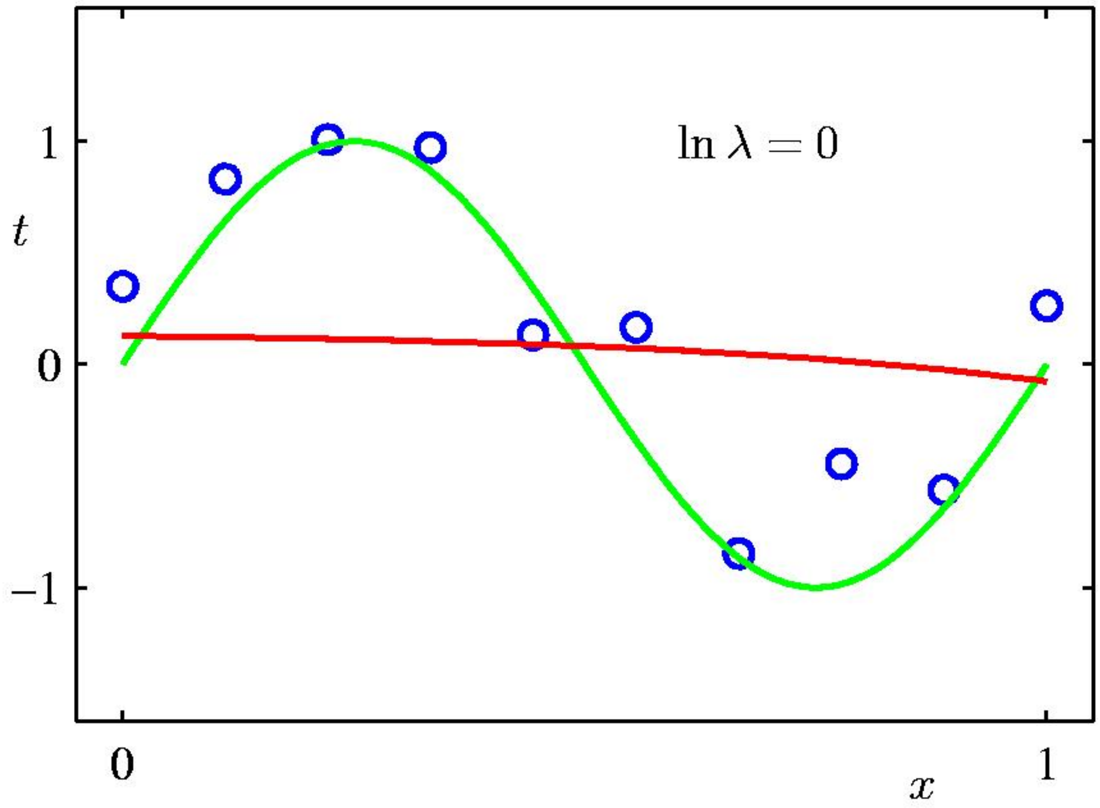


Figure D.2. A linear model that fits the data given in Figure D.1.

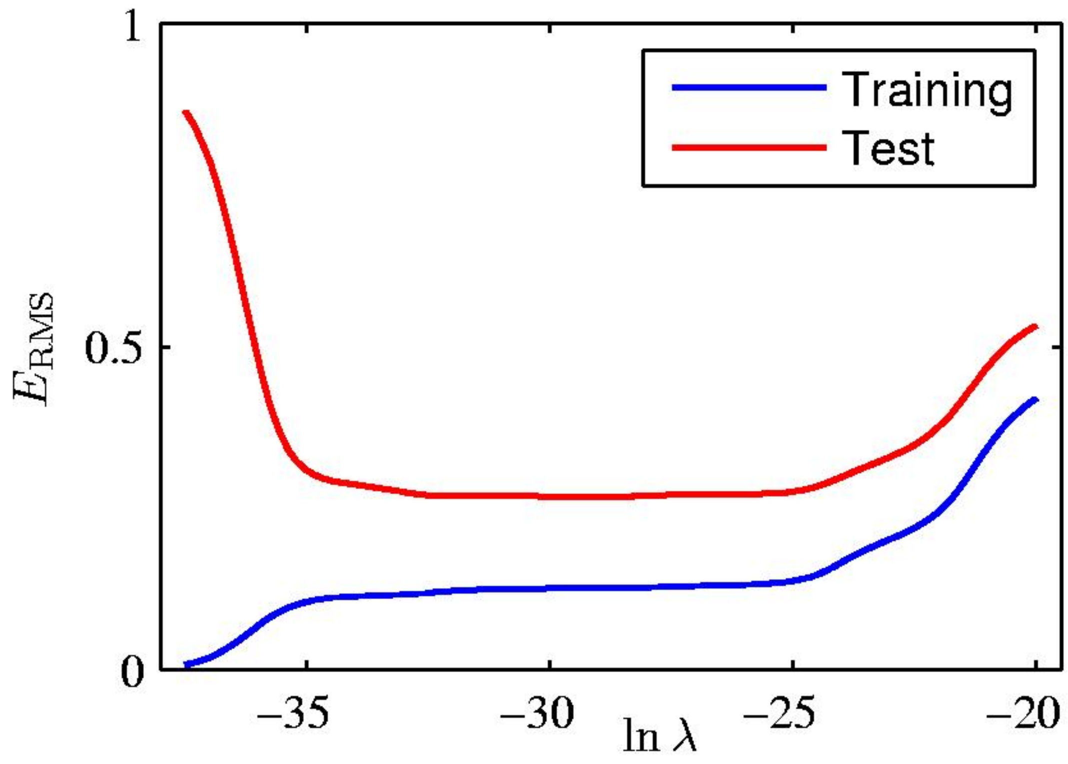
Regularization: $\ln \lambda = -18$



Regularization: $\ln \lambda = 0$



Regularization: E_{RMS} vs. $\ln \lambda$

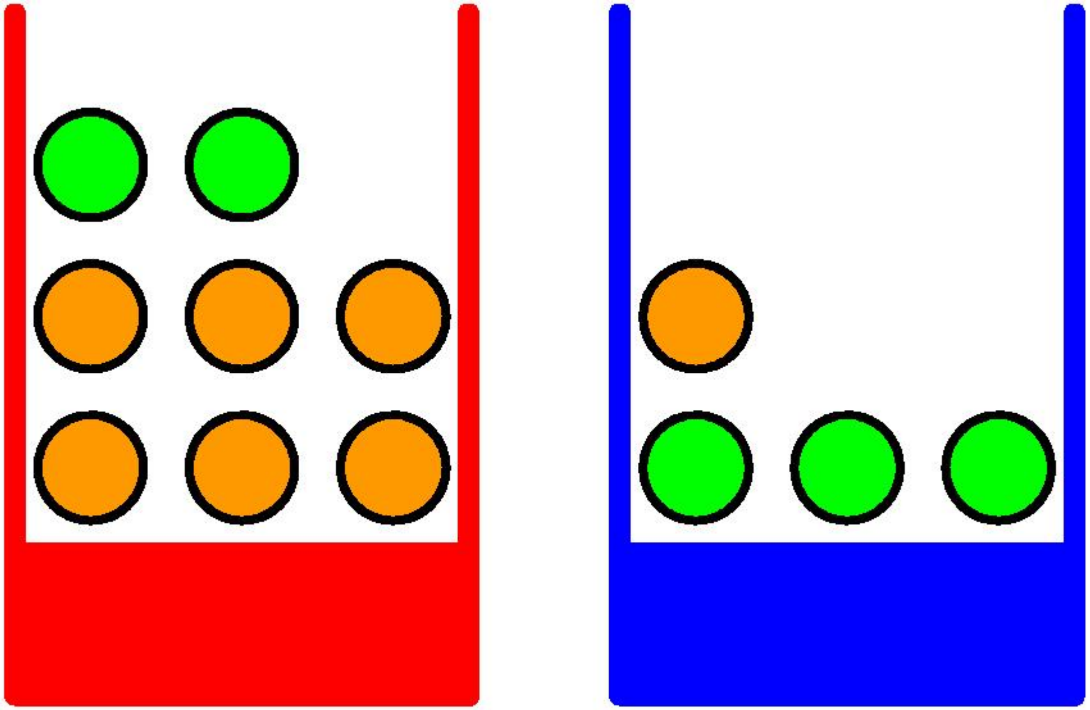


Polynomial Coefficients

	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

Probability Theory

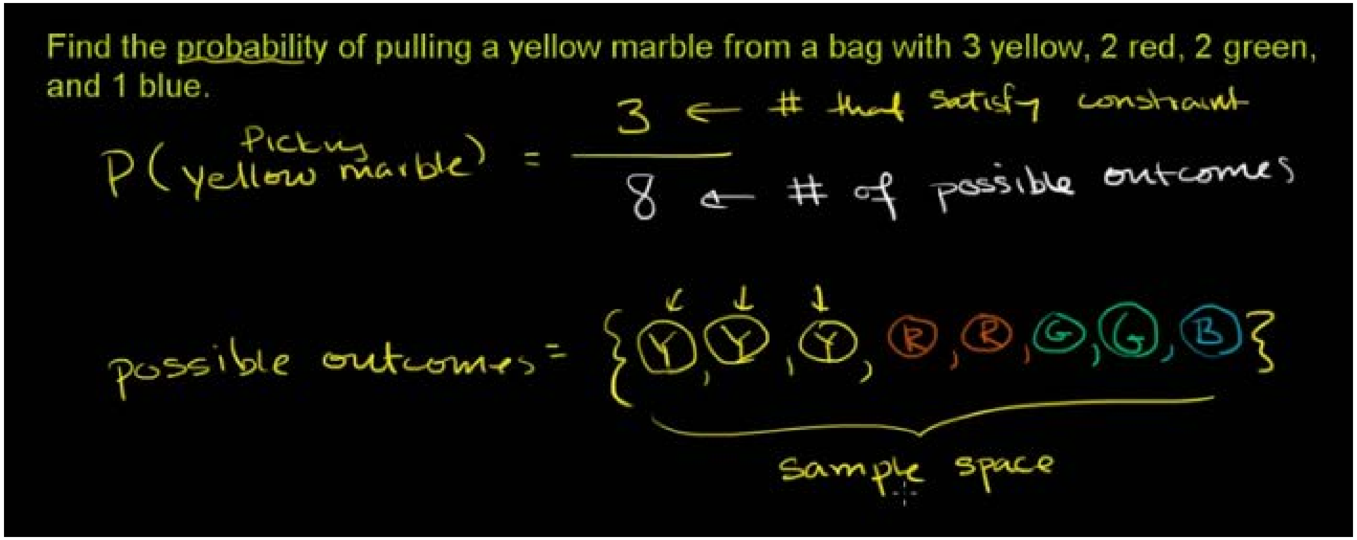
Apples and Oranges

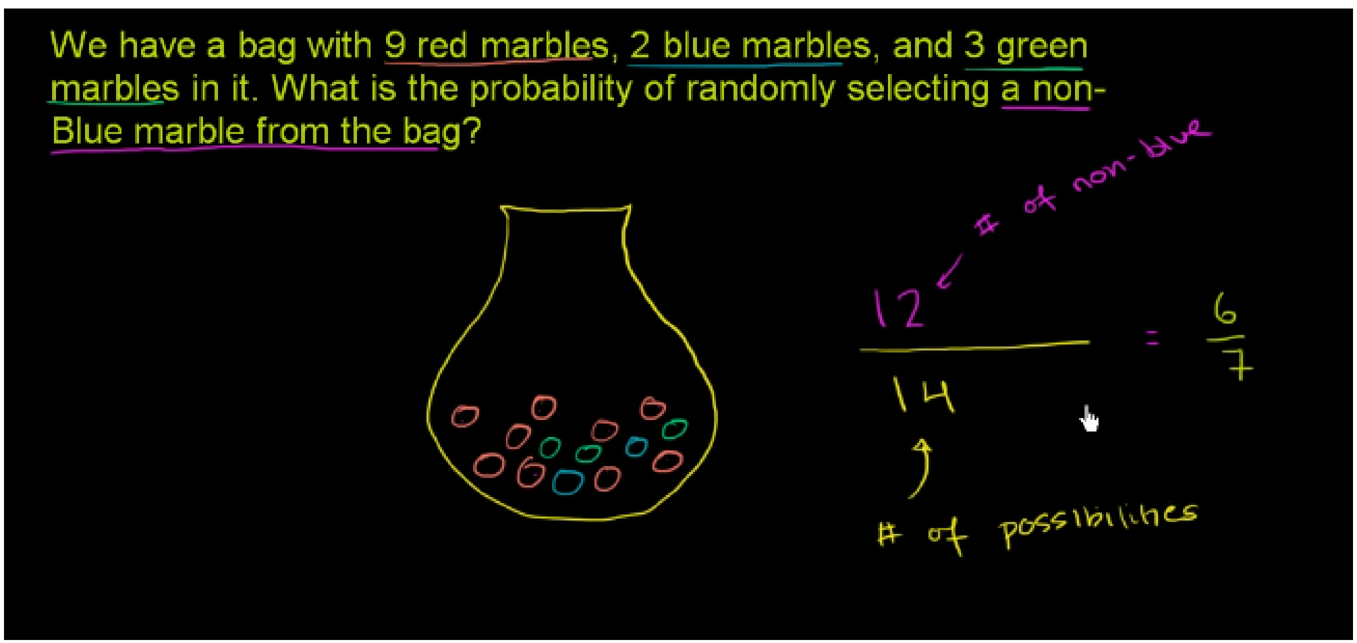


From Bernoulli to Multinomial Distribution

- A brief review of probability, Bernoulli distribution, binomial distribution and multinomial distribution.
- Multinomial distribution plays vital important role in data mining/machine learning:
  - The basic model of English text, documents, fundamental theory for information retrieval, search engine, etc.
  - The basis for logistic regression and neural networks.
- An alternative (better) model of Naïve Bayes Classification

Probability = count possible outcomes satisfying requirements/constraints





Bernoulli distribution: Simplest probability distribution

Today is sunny or not-sunny.
Your team win or lose.
You throw a coin; it is head-up or head-down
Yow throw a die; the result is 6, or it is not 6 (which is 1 or 2 or 3 or 4 or 5)

Bernoulli Distribution

A **Bernoulli distribution** arises from a random experiment which can give rise to just two possible outcomes. These outcomes are usually labeled as either "success" or "failure." If p denotes the probability of a success and the probability of a failure is (1 - p), the the Bernoulli probability function is

$$P(0) = (1 - p) \quad and\ P(1) = p$$

Binomial Distribution :

Y = X1 + … + Xn : sum of N independently identically distributed Bernoulli random variables

One experiment:

- the experiment consists of n independent trials, each with two mutually exclusive outcomes (**success** and **failure**)
- for each trial the probability of success is p (and so the probability of failure is $1 - p$)

Each such trial is called a **Bernoulli trial**.

Experiment: Throwing N identical coins, head-up/head-down

Experiment: Throwing one coin N times

Binomial Distribution Formula

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

where

n = the number of trials (or the number being sampled)

x = the number of successes desired

p = probability of getting a success in one trial

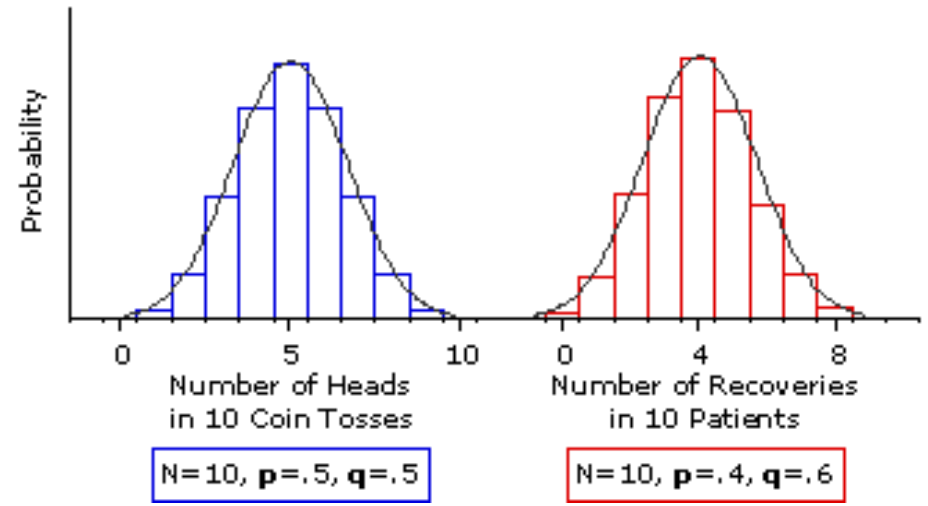
$q = 1 - p$ = the probability of getting a failure in one trial

Example 1

Q. A coin is tossed 10 times. What is the probability of getting exactly 6 heads?

$$p = 0.5,\ q = 1 - p = 0.5,\ n = 10,\ x = 6$$

$$P(x=6) = \binom{10}{6} 0.5^6\ 0.5^{(10-6)} = 0.2051$$
$$P(x=5) = 0.2461$$
$$P(x=3) = P(x=7) = 0.1172$$
$$P(x=2) = P(x=8) = 0.0439$$



Example 3

60% of people who purchase sports cars are men. If 10 sports car owners are randomly selected, find the probability that exactly 7 are men.

$$p = 0.6,\ q = 1 - p = 0.4,\ n = 10,\ x = 7$$

$$P = \binom{10}{7} 0.6^7\ 0.4^{(10-7)} = 0.215$$

Multinomial Distribution

- The Binomial distribution can be extended to describe number of outcomes in a series of independent trials each having more than 2 possible outcomes.

- If a given trail can result in the k outcomes $E_1, E_2, \ldots, E_k$ with probabilities $p_1, p_2, \ldots, p_k$, then the probability distribution of the random variables $X_1, X_2, \ldots, X_k$, representing the number of occurrences for $E_1, E_2, \ldots, E_k$ in n independent trials is

$$p_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

with $\sum_{i=1}^{k} x_i = n$, and $\sum_{i=1}^{k} p_i = 1.$

Example:

The distribution of blood types in the US is:

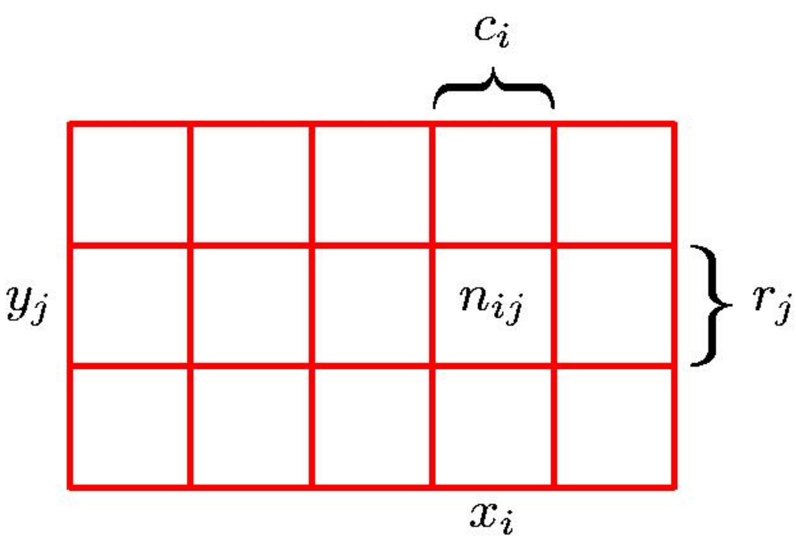
Type	O	A	B	AB
Probability	0.44	0.42	0.10	0.04

In a random sample of 10 Americans, what is the probability 6 have blood type O, 2 have type A, 1 has type B, and 1 has type AB?

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

$$P(X_1=6, X_2=2, X_3=1, X_4=1) = \frac{10!}{6!2!1!1!} 0.44^6 \, 0.42^2 \, 0.10^1 \, 0.04^1 = 0.01290$$

Probability Theory



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

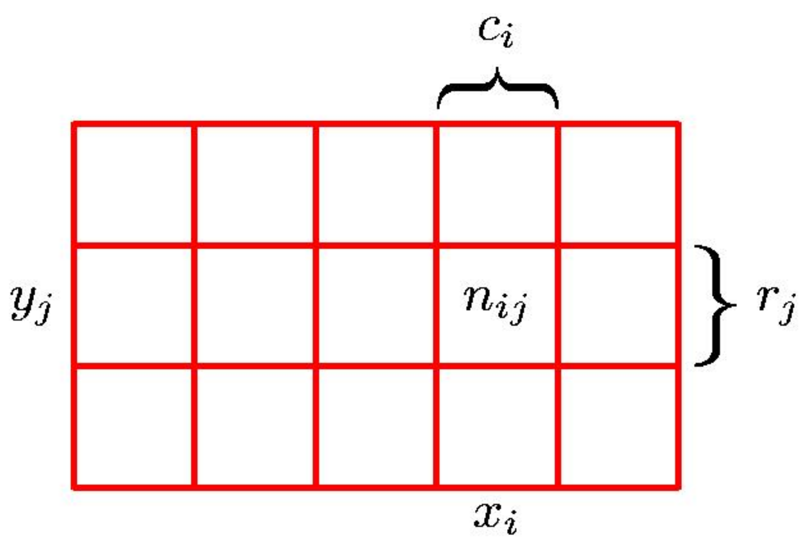
Joint Probability

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

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Probability Theory



Sum Rule

$$
\begin{aligned}
p(X = x_i) = \frac{c_i}{N} &= \frac{1}{N}\sum_{j=1}^{L} n_{ij} \\
&= \sum_{j=1}^{L} p(X = x_i, Y = y_j)
\end{aligned}
$$

Product Rule

$$
\begin{aligned}
p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\
&= p(Y = y_j | X = x_i) p(X = x_i)
\end{aligned}
$$

The Rules of Probability

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

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

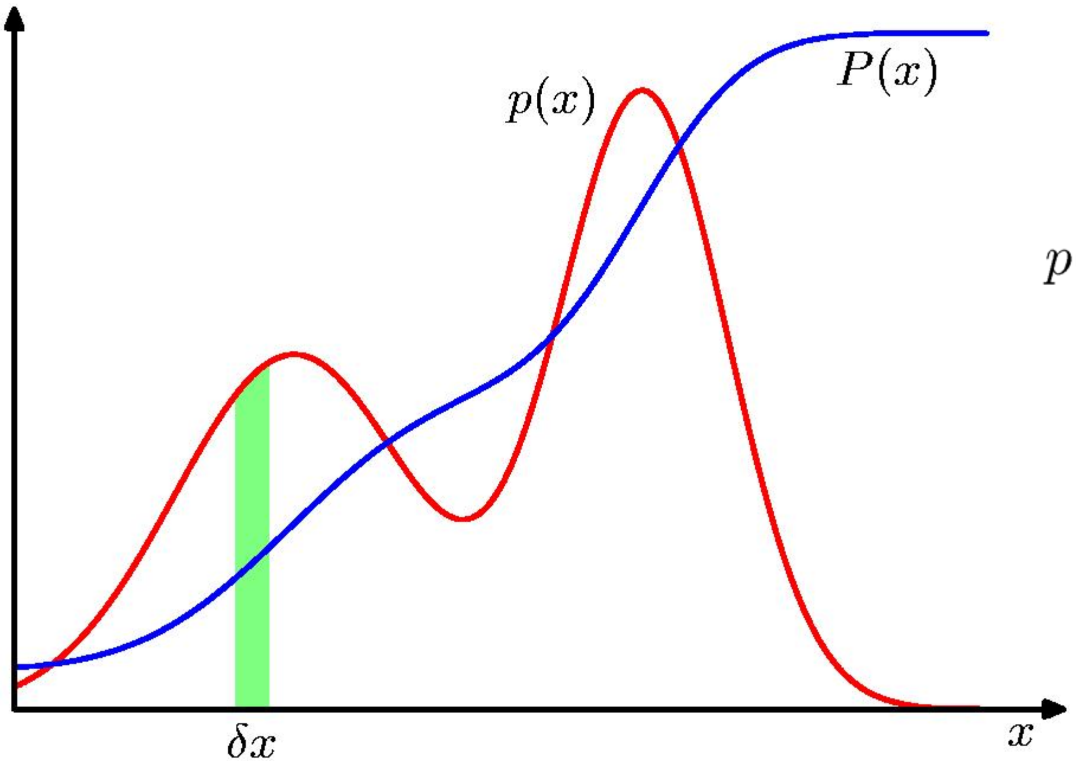
Bayes' Theorem

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

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

posterior $\propto$ likelihood × prior

Probability Densities

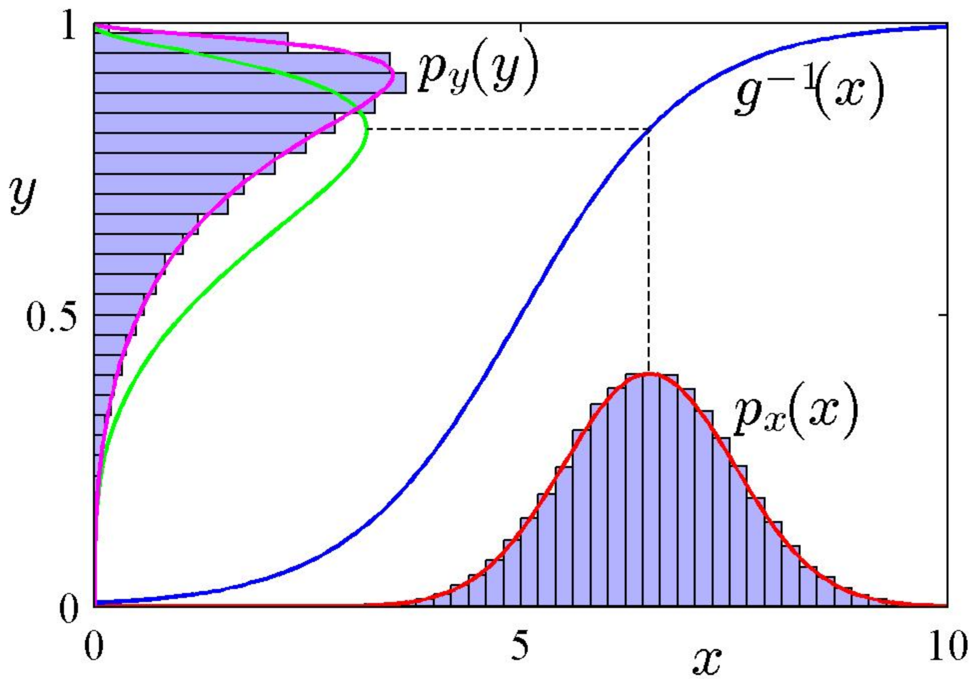


$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Transformed Densities



$$
\begin{aligned}
p_y(y) &= p_x(x) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| \\
&= p_x(g(y)) \, |g'(y)|
\end{aligned}
$$

Expectations

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x$$

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Conditional Expectation
(discrete)

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

Approximate Expectation
(discrete and continuous)

Variances and Covariances

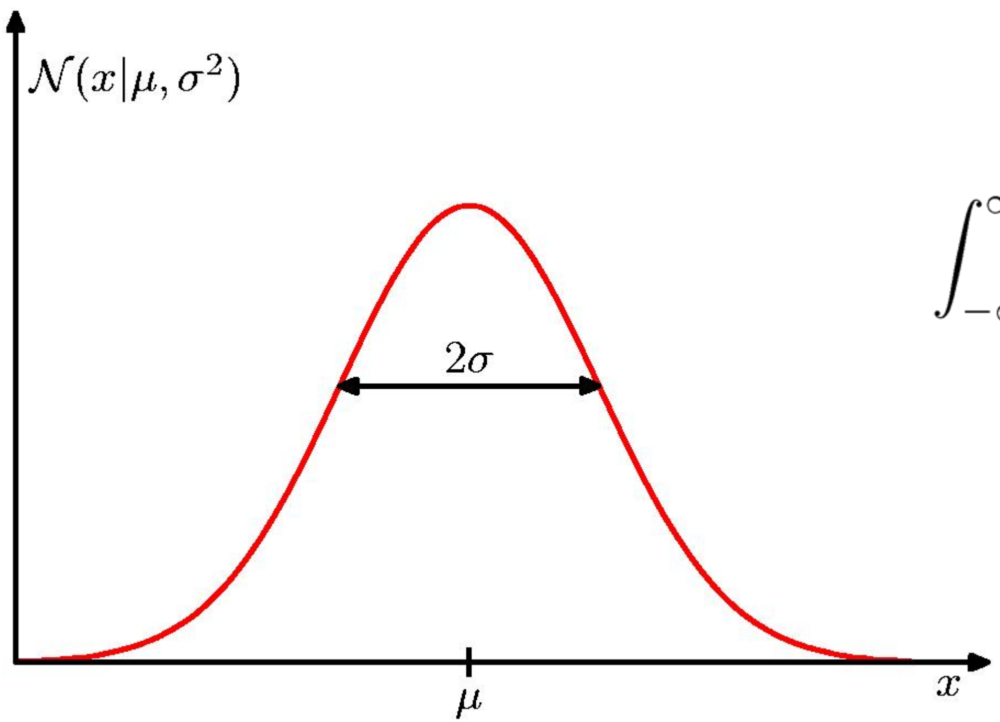
$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

$$\begin{aligned}
\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]
\end{aligned}$$

$$\begin{aligned}
\mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]
\end{aligned}$$

The Gaussian Distribution

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



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) \mathrm{d}x = 1$$

Gaussian Mean and Variance

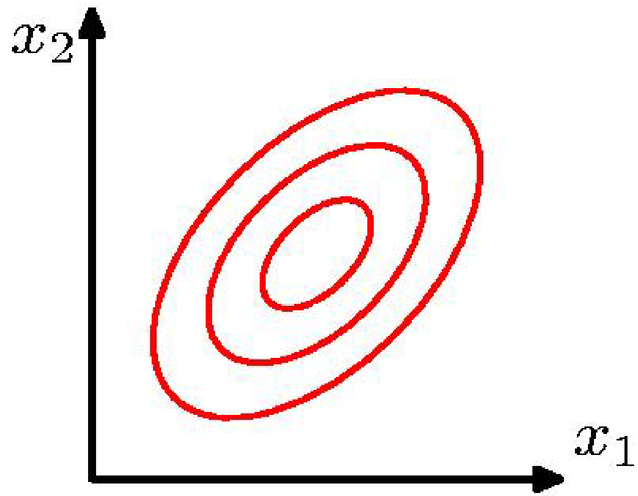
$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \, \mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

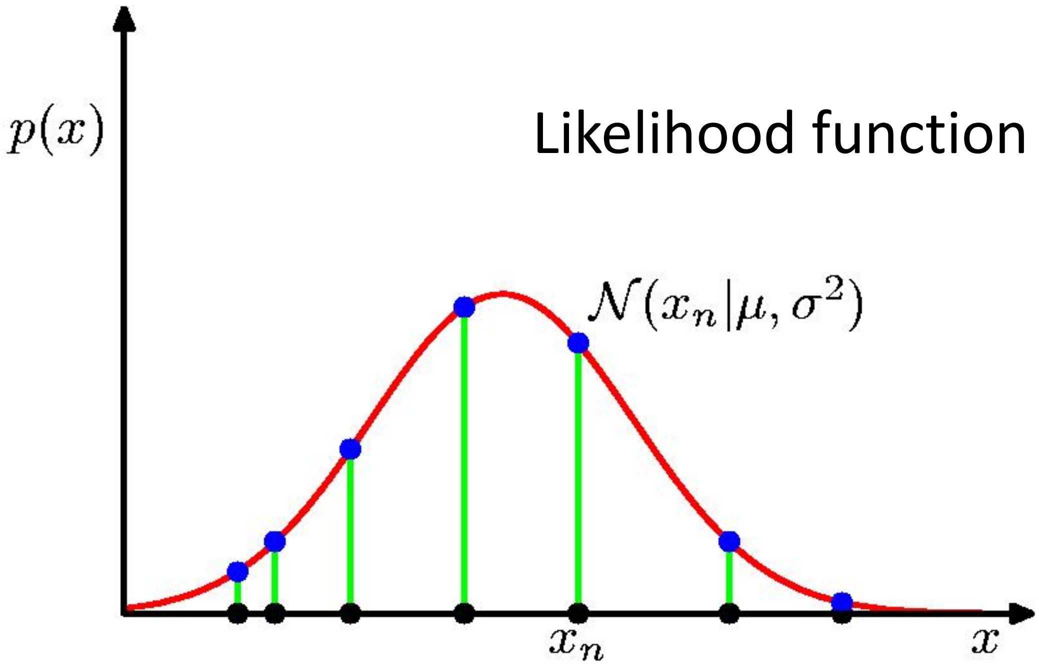
$$\operatorname{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

The Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$



Gaussian Parameter Estimation



$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu, \sigma^2\right)$$

Maximum (Log) Likelihood

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

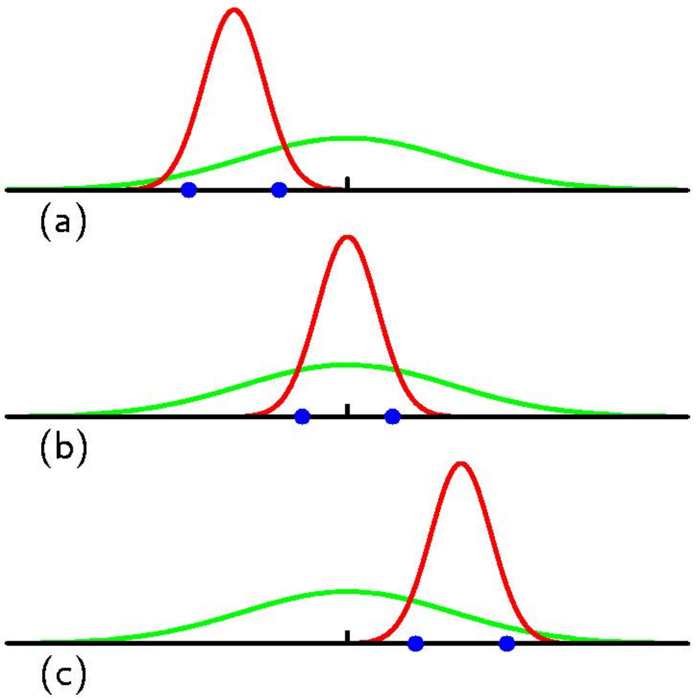
$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

Properties of μ_{ML} and σ^2_{ML}

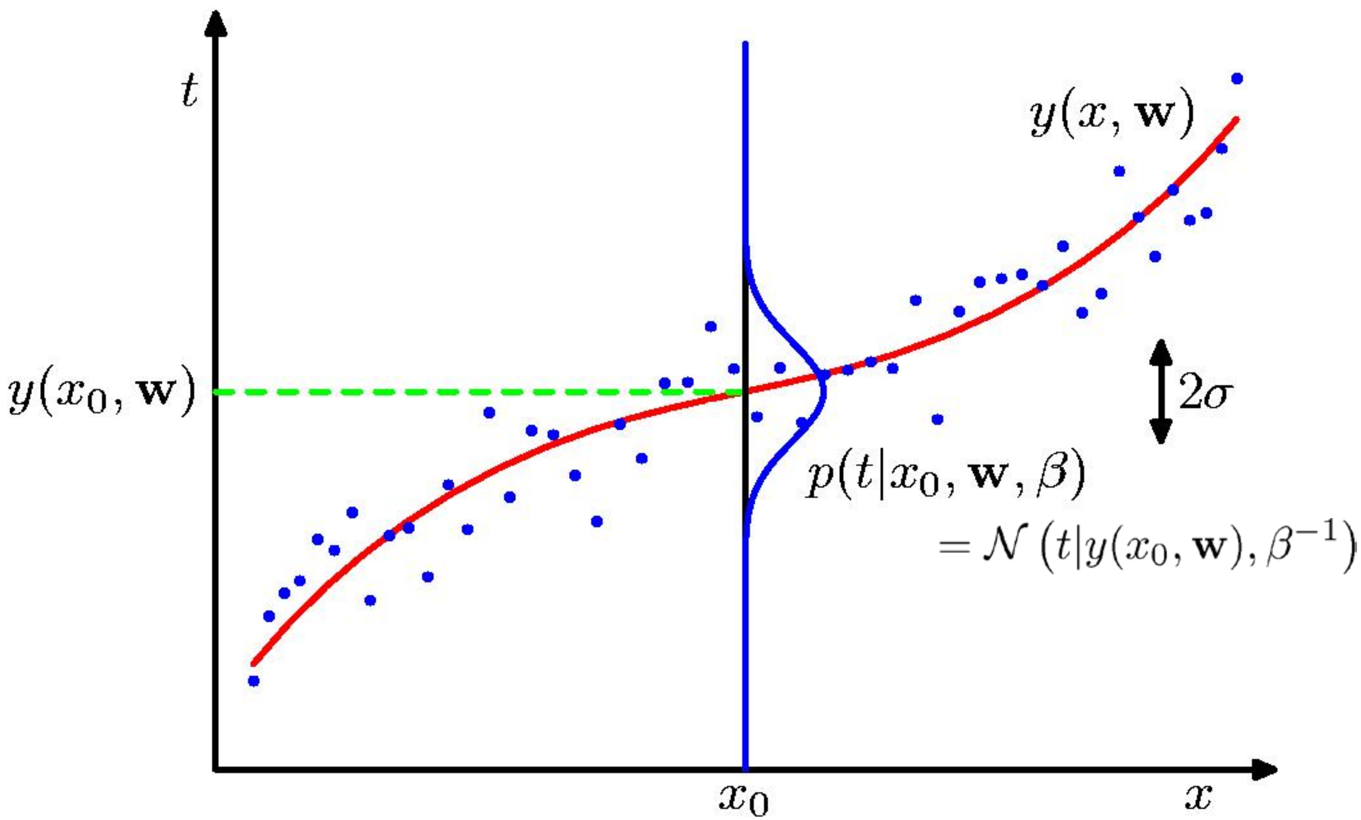
$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2$$

$$\begin{aligned}\widetilde{\sigma}^2 &= \frac{N}{N-1}\sigma^2_{\mathrm{ML}} \\ &= \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2\end{aligned}$$



Curve Fitting Re-visited



Maximum Likelihood

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

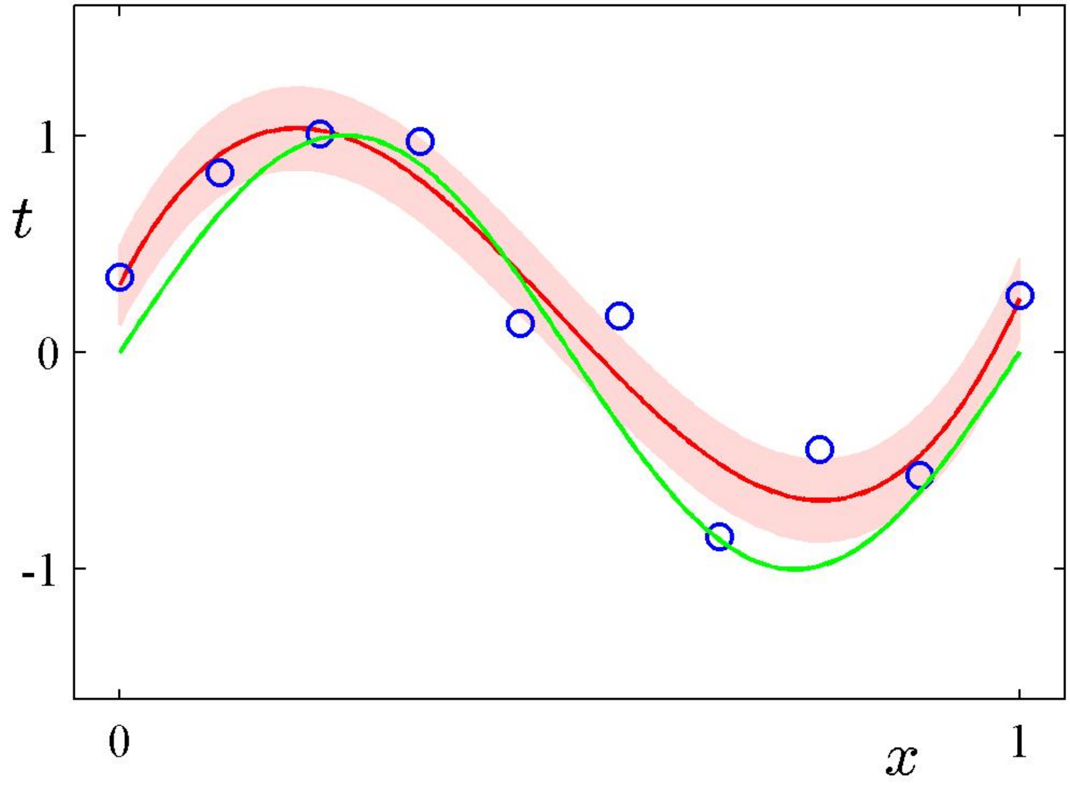
$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \underbrace{-\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2}_{\beta E(\mathbf{w})} + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

Determine $\mathbf{w}_{\mathrm{ML}}$ by minimizing sum-of-squares error, $E(\mathbf{w})$.

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2$$

Predictive Distribution

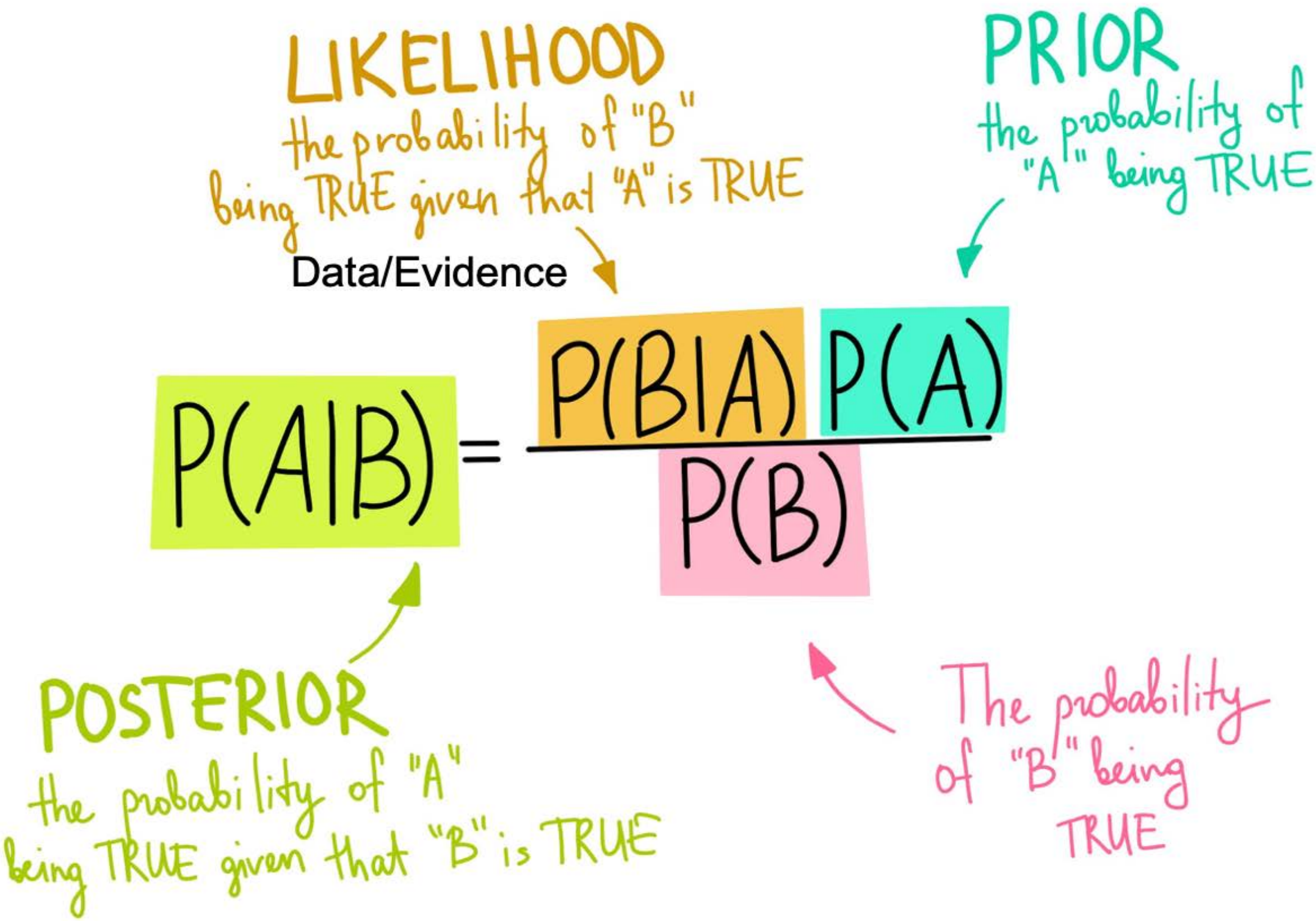
$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right)$$



Bayes Classification

Using Bayes Theorem (conditional probability) to obtain the class/label posterior probability of a data instance given its observed data (attributes/features)

Reading: Textbook Sections 5.3, 5.3.1, 5.3.2, 5.3.3



This formula is useful ONLY when A is class/hypothesis

Example: Test of Viral Infection

A medical test for a viral infection. It is 95% reliable for infected patients and 99% reliable for the healthy ones：

If a patient has the virus (event V), and the test shows that (event S) with probability

$$P\{S \mid V\} = 0.95$$

If a patient does not have the virus, the test confirms that with probability

$$P\{\bar{S} \mid \bar{V}\} = 0.99$$

A patient tests positive (the test shows that the patient has the virus).

Does this means he has 95% probability of the virus?

No!

Because the question refers the probability that he has the virus and the test confirms that, i.e., P{V|S} . This quantity is not given directly in the statement of the problem.

We compute P{V|S} using Bayes theorem.

Bayes' Rule

Bayes Theorem (conditional probability):

$$\boxed{P\{B \mid A\} = \frac{P\{A \mid B\}P\{B\}}{P\{A\}} = \frac{P\{A \mid B\}P\{B\}}{P\{A \mid B\}P\{B\} + P\{A \mid \overline{B}\}P\{\overline{B}\}}}$$

Law of Total Probability

$$P\{A\} = \sum_{j=1}^{k} p\{\mathrm{A} \mid B_j\} P\{B_j\}$$

In case of two events (k=2),

$$P\{A\} = P\{A \mid B\}P\{B\} + P\{A \mid \overline{B}\}P\{\overline{B}\}$$

Medical Test Example cont.

~~We need additional information:~~ Suppose 4% of all the population are infected with the virus, P{V} = 0.04.

Recall: $P\{S \mid V\} = 0.95 \quad P\{\bar{S} \mid \bar{V}\} = 0.99$

The desired (conditional) probability is

$$P\{V \mid S\} = \frac{P\{S \mid V\}P\{V\}}{P\{S \mid V\}P\{V\} + P\{S \mid \bar{V}\}P\{\bar{V}\}}$$

$$= \frac{(0.95)(0.04)}{(0.95)(0.04) + (1 - 0.99)(1 - 0.04)} = 0.7983$$

Test of Viral Infection - Conclusion

Thus the probability of the patient has the virus is 79.83%, not 95%.

Bayesian view:

This patient has 4% probability of been infected by the virus [because 4% of the population has the virus]. Because now he tested positive for the virus, his chance of virus increased to 79.83%.

This patient has 4% probability of been infected by the virus [because 4% of the population has the virus (prior probability)]. Because now he tested positive for the virus (new data evidence), his chance of virus increased to 79.83%.

Naïve Bayes Classification

Using Bayes Theorem (conditional probability) to obtain the class/label posterior probability of a data instance given its observed data (attributes/features)

5.3.3 Naïve Bayes Classifier

A naïve Bayes classifier estimates the class-conditional probability by assuming that the attributes are conditionally independent, given the class label y. The conditional independence assumption can be formally stated as follows:

$$P(\mathbf{X}|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y), \tag{5.12}$$

where each attribute set $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ consists of d attributes.

Probability of occurrence of X is equal to the product of the probability of occurrence of every attributes of X given the class of X

This says each class has a different multinomial distribution of attributes.

To classify a test record, the naïve Bayes classifier computes the posterior probability for each class Y:

$$P(Y|\mathbf{X}) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y)}{P(\mathbf{X})}. \tag{5.15}$$

Since $P(\mathbf{X})$ is fixed for every Y, it is sufficient to choose the class that maximizes the numerator term, $P(Y)\prod_{i=1}^{d} P(X_i|Y)$.

the prior probability $P(Y)$

the class-conditional probabilities $\prod_i P(X_i|Y)$. = multinomial distribution of attributes for class Y

Compute probability of occurrence of each attributes for class Y="no"

Compute probability of occurrence of each attributes for class Y="yes"

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(a)

P(Home Owner=Yes|No) = 3/7
P(Home Owner=No|No) = 4/7
P(Home Owner=Yes|Yes) = 0
P(Home Owner=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No) = 1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/3
P(Marital Status=Divorced|Yes) = 1/3
P(Marital Status=Married|Yes) = 0

For Annual Income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

(b)

Figure 5.10. The naïve Bayes classifier for the loan classification problem.

Standard multinomial distribution parameter estimation:

$$P(x_i = n_i | Y = y)^{\mathrm{MLE}} = p_{i,y}^{\mathrm{MLE}} = \frac{n_{i,y}}{N_y}$$

where $n_{i,y}$ is the number of training examples in class y where attribute x_i occurs, N_y is the number of training examples in class y.

Laplace smoothed multinomial distribution parameter estimation:

See 2nd Edition Textbook p.224

$$P(x_i = n_i | Y = y)^{\mathrm{smoothed}} = p_{i,y}^{\mathrm{smoothed}} = \frac{n_{i,y} + 1}{N_y + \nu_y}$$

where ν is the total number of times attribute x_i occurs in class y training examples.

In most applications, we use Laplace smoothed parameter estimation