CSE 574/474 Midterm Study Guide

Fall 2015

Contents

1 Midterm Overview

Time and Place

- Monday, October 19^{th}, 2015
- $6{:}30_{PM}$ – $7{:}50_{PM}$, Hochstetter 114

Topic and Format

- Topics: Probability, Linear Regression, Linear Classification, Neural Network
- Closed Book, closed notes. No calculators.
- The exam is worth 20% of the final grade.

Textbook Sections

- Ch.1, 1.1 and 1.2
- Ch.2, 2.1, 2.3.1 – 2.3.6 and 2.4.2
- Ch.3, 3.1, 3.2 and 3.3
- Ch.4, 4.1.1 – 4.1.3, 4.1.7, 4.2.1 – 4.2.3 and 4.3.1 – 4.3.2
- Ch.5, 5.1, 5.2, and 5.3.1 – 5.3.2

Specific Topics

- **Probability Theory**
  - Be able to compute conditional, marginal, and joint probability.
  - Be able to compute mean, variance, covariance and entropy.
  - Understand Bayes' theorem, include Sum Rule and Product Rule.
- **Probability Densities**
  - At least know the density functions of Bernoulli and Gaussian distributions.
  - Know the definition of *conjugate prior*, be able to identify whether a pair of likelihood distributions and prior distributions has the property of conjugacy for Bayesian Analysis
- **ML, MAP Estimation**
  - Know the gradient descent solution for parameter estimations.
  - Be able to write down likelihood/prior/posterior function for given data and assumption of density distribution.
  - Be able to derive maximum likelihood estimations (ML) and maximum *a posterior* estimation (MAP) of parameters.
- **Linear Models for Linear Regression**
  - Know error function, basis function and design matrix.
  - Understand how regularization works.
  - Model complexity vs. performance.
  - Know what over-fitting is, and possible ways to overcome it.
  - Know the equivalence of the least-square method and maximum likelihood estimation under Gaussian noised model.
  - *Bayesian Method*: At least know the advantages of the Bayesian method over Point Estimation methods (ML, MAP).

- **Linear Models for Classification**
  - Understand least square method and know under what circumstances it could fail.
  - Know what generative models and discriminative models are, and the difference between them.
  - Know what the perceptron algorithm is and how to do the simple calculation.
  - *Logistic Regression*: Be able to write down the error function, and learning rule of updating weights in gradient descent manner.
- **Neural Networks**
  - Feed-forward Network Functions, Network Training, and error backpropagation.

2 Notes on Covered Topics

2.1 Notations and Errata

A few things are important, that don't fall into other sections.

A **Probability Distribution Function** has an ambiguous definition. It can refer to:

- **Discrete Case:** Probability Mass Function (PMF)
- **Continuous Case:** Probability Density Function (PDF)
- **Both Cases:** Cumulative Distribution Function (CDF)

In this guide and throughout the class, **PDF** refers to either the PDF or the PMF, whereas CDF is always explicitly stated.

The CDF and the PDF are intimately related. It is known:

$$PDF(x) = \frac{\partial}{\partial X} CDF(x) \tag{1}$$

$$CDF(x) = \int_{-\infty}^{x} PDF_x(t)dt \tag{2}$$

The Greek letter Phi means three different things.

$$\phi(x) = \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}} = \mathcal{N}(0,1) \tag{3}$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \tag{4}$$

$$\varphi(x) = E[e^{itX}] \text{ where } t \in \mathbb{R} \text{ is the argument} \tag{5}$$

Likelihood and **Probability** are closely related, but not the same. The likelihood of a set of parameter values θ is equal to the probability of those observed outcomes given those parameters values, i.e.,

$$\mathcal{L}(\theta|x) = p(x|\theta) \tag{6}$$

$$\mathcal{L}(\theta|x) = p_\theta(x) = P_\theta(X = x) = P(X = x; \theta) \text{ in a discrete PDF} \tag{7}$$

$$\mathcal{L}(\theta|x) = f_\theta(x) \text{ in a continuous PDF} \tag{8}$$

w^T represents a column vector of coefficients w. A simple way to remember it is:

$$y(x, w) = \sum_{j=0}^{M-1} w_j \phi_j(x) = w^T \phi(x) \tag{9}$$

The **sigmoid** function maps real $a \epsilon (-\infty, \infty)$ to a finite $(0, 1)$ interval, and is defined as:

$$\sigma(a) = \frac{1}{1 + e^{-a}} \tag{10}$$

2.2 Probability Theory

2.2.1 . . . be able to compute conditional, marginal, and joint probability.

Conditional Probability is the probability of an event (A), given that another event (B) has occurred. Written as $p(A|B)$, and the *Kolmogorov definition* is:

$$p(A|B) = \frac{p(A \cap B)}{p(B)} \tag{11}$$

If A and B are independent, then the probability of each event happening is not dependent on the other. In other words, $p(A|B) = p(A)$.

Marginal Probability is closely related to **Joint Probability**. If we know the joint distribution of two events A and B, the *marginal probability* of an event A is the probability that event A happens when event B is not known. The *joint probability* is the probability that **both** events occur.

As an example of marginal, joint, and conditional, assume that we can map out the possibility of a pedestrian being hit by a car at an intersection based upon the status of the stop light. Take $H = \{0, 1\}$, with 0 being *not hit*, and 1 being *hit*, and take $L = \{R, Y, G\}$ as the color of the stop light. An example of a conditional probability table would be:

$P(H\|L)$			
	L=R	L=Y	L=G
H=0	0.99	0.9	0.2
H=1	0.01	0.1	0.8

If we know $p(L = l)$, then we can map the *joint* probability. If $p(L = R) = 0.2, p(L = Y) = 0.1, p(L = G) = 0.7$:

Joint Distribution: $p(H, L)$				
	L=Red	L=Yellow	L=Green	Marginal Probability $p(H)$
H=0	0.198	0.09	0.14	0.427
H=1	0.002	0.01	0.56	0.572
Total	0.2	0.1	0.7	1

2.2.2 . . . be able to compute mean, variance, covariance, and entropy.

Mean is denoted as μ. In the *absolute* sense, as in "the average value of a dice throw", we would calculate it as:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{12}$$

Where we have n observations, and x_i is the i^{th} observation. In the probability sense, we talk about **Expected Value**. To calculate this, we multiply an observation x_i by its associated probability $p_i = p(X = x_i)$, and get:

$$E[x] = \sum_{i=1}^{n} x_i p_i \tag{13}$$

If we have a *continuous random variable*, with a probability distribution of $f(x)$, then it can be computed as:

$$E[x] = \int_{-\infty}^{\infty} x f(x) dx \tag{14}$$

Variance measures how much variability there is in $f(x)$ around its mean (or *expected*) value $E[f(x)]$. The variance of $f(x)$ can be calculated as:

$$var[f] = E[(f(x) - E[f(x)])^2] = E[f(x)^2] - E[f(x)]^2 \tag{15}$$

Similarly, the variance of the variable x itself is:

$$var[x] = E[x^2] - E[x]^2 \tag{16}$$

For two random variables x and y, their **covariance** is defined as:

$$\begin{aligned} cov[x, y] &= E_{(x,y)}[\{x - E[x]\}\{y - E[y]\}] \\ &= E_{(x,y)}[xy] - E[x]E[y] \end{aligned}$$

This expresses how x and y vary together. If x and y are independent, then their covariance vanishes — similarly, if $\vec{x}$ and $\vec{y}$ are vectors of random variable, then their covariance is a matrix.

Entropy is a measure of the *unpredictability* of a probability mass function. In a high level, it is written as:

$$H(X) = E[I(X)] = E[-\ln P(X))]. \tag{17}$$

$I(X)$ is the *information content* of X, which is the negative log of the probability mass function $p(X)$. In the event of X being a discrete random variable, we can use the definition of the expected value operator and get:

$$H(X) = \sum_{i=0}^{n} P(x_i) I(x_i) = -\sum_{i=0}^{n} P(x_i) log_b P(x_i) \tag{18}$$

Where b is the base of the logarithm, generally e. If $p(X)$ is a continuous distribution, we can again use the value of $E[x]$:

$$H(X) = \int P(x) I(x) dx = -\int P(x) log_b P(x) dx \tag{19}$$

2.2.3 . . . understand Bayes' Theorem, including Sum Rule and Product Rule.

Bayes' Theorem states:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \text{ where } p(X) = \sum_Y p(X|Y)P(Y) \tag{20}$$

The Sum Rule of Probability Theory states that, when considering two random variables, with N trials sampling both X and Y, then taking $n_{(i,j)}$ which represents the number of trials where $X = x_i$ and $Y = y_j$, and N which represents the total trials, the joint probability is:

$$P(X = x_i, Y = y_j) = \frac{n_{(i,j)}}{N} \tag{21}$$

Similarly, the marginal probability is:

$$p(X = x_i) = \frac{c_i}{N}, \text{ where } c_i = \sum_j n_{(i,j)} \text{ or } P(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j) \tag{22}$$

Put into English, given a joint probability of two variables X and Y, the probability that $X = x_i$ is the sum of all probabilities where $X = x_i$, regardless of Y. **The Product Rule** allows us to decompose joint probabilities. See two examples below:

$$p(X = x_i, Y = y_j) = p(X = y_i|Y = y_j)P(Y = y_j) \tag{23}$$

$$p(X = x_i, Y = y_j, Z = z_k) = p(X = x_i, Y = y_j|Z = Z_k)p(Z = z_k) \tag{24}$$

$$= p(X = x_i|Y = y_j, Z = z_k)p(Y = y_j|Z = z_k)p(Z = z_k) \tag{25}$$

2.3 Probability Densities

2.3.1 . . . at least know the density functions of Bernoulli and Gaussian distributions

The probability density function for a **bernoulli distribution** is simple:

$$p(x = k|\mu) = \begin{cases} \mu & \text{if } k = 1 \\ 1 - \mu & \text{if } k = 0 \end{cases} \tag{26}$$

The **joint probability density function of a bernoulli distribution** is:

$$p(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1 - \mu)^{1-x_n} \tag{27}$$

This follows directly from the definition - we are simply multiplying all of their probabilities.

The **Gaussian Distribution**, or **Normal Distribution**, takes the form:

$$\mathcal{N}(X|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} \tag{28}$$

With arguments μ as the mean and σ^2 as the variance. The other parts worth noting is the *standard deviation* $\sqrt{\sigma^2} = \sigma$, and the precision $\beta = \frac{1}{\sigma^2}$.

2.3.2 . . . know the definition of *conjugate prior*, be able to identify whether a pair of likelihood distributions and prior distributions has the property of conjugacy for Bayesian Analysis

If a posterior distribution $p(\theta|x)$ is in the same family as a prior distribution $p(\theta)$, then they are called **conjugate distributions.** As an example, **Gaussian** distributions are said to be self-conjugate, meaning that if we choose a prior distribution that is Gaussian, the posterior distribution will then also be Gaussian. The other distributions we need to know are:

$$\text{Multinomial} \Leftrightarrow \text{Dirichlet}$$
$$\text{Binomial} \Leftrightarrow \text{Beta}$$

2.4 ML, MAP Estimation

2.4.1 . . . know the gradient descent solution for parameter estimations

We will use the following error function in describing gradient descent:

$$E_D(w) = \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - w^T\varphi(x_n)\right\}^2 \tag{29}$$

If we denote $E_D(w) = \sum_n E_n$, we can update the parameter vector $\vec{w}$ using:

$$w^{(\tau+1)} = w^{(\tau)} - \eta\nabla E_n \tag{30}$$

And, substituting for the derivative, we get:

$$w^{(\tau+1)} = w^{(\tau)} + \eta(t_n - w^{(\tau)}\phi(x_n))\phi(x_n) \tag{31}$$

Where w is initialized to some starting vector $w^{(0)}$, and η is chosen to ensure convergence. The gradient vector ∇ is defined as:

$$\nabla E[\vec{w}] \equiv \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \ldots, \frac{\partial E}{\partial w_n}\right] \tag{32}$$

The **negation** of this vector specifies the direction of steepest decrease. The gradient descent rule is then:

$$\vec{w} \leftarrow \vec{w} + \Delta\vec{w} \text{ where } \Delta\vec{w} = -\eta\nabla E[\vec{w}] \tag{33}$$

With η as some positive constant, called the **learning rate**.

2.4.2 . . . be able to write down likelihood/prior/posterior function for given data and assumption of density distribution

To this end, we make a **Bayesian Approach**. Remember:

$$posterior\ \alpha\ likelihood \times prior \tag{34}$$

Which is **Bayes' Rule** in words. We will first assume a **prior** density distribution as a **multivariate Gaussian** for w, in other words,

$$p(w) = \mathcal{N}(w|m_0, S_0) \tag{35}$$

With mean m_0, and covariance matrix S_0. For a **likelihood**, we assume a noise precision parameter β, so the likelihood $p(t|w)$ with Gaussian noise has the exponential form:

$$p(t|X, w, \beta) = \mathcal{N}\big(t|y(x, w), \beta^{-1}\big) = \prod_{n=1}^{N} \mathcal{N}\Big(t_n|w^T\phi(x_n), \beta^{-1}\Big) \tag{36}$$

Finally, if we have **prior** $p(w)$ and **likelihood** $p(t|w)$, we need **posterior** $p(w|t)$:

$$p(w|t) = \mathcal{N}(w|m_N, S_N) \tag{37}$$

$$m_N = S_n(S_0^{-1}m_0 + \beta\Phi^T t) \tag{38}$$

$$S_N^{-1} = S_0^{-1} + \beta\Phi^T\Phi \tag{39}$$

The **Maximum Posterior Weight** Vector (MAP) is $w_{MAP} = m_N$.

2.4.3 . . . be able to derive maximum likelihood estimations (ML) and maximum *a posterior* estimation (MAP) of parameters

Remember that maximizing likelihood is equivalent to minimizing error. The error function:

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}\big\{y(x_n, w) - t_n\big\}^2 \tag{40}$$

The error function can take on many forms. First, $y(x_n, w)$ and t_n can have their order changed - as we're measuring the **distance**, and we're squaring it, it is irrelevant. t_n represents the *target* value, in other words, a **known** value, given some x_n. $y(x_n, w)$ is merely the PDF describing the model, so it may be abbreviated as $w \cdot f(x)$, and w may be written as w^*, denoting the minimum error coefficient, or w^T, denoting a column vector of coefficients.

As described in later sections, we take the derivative of this function and set equal to zero, giving us:

$$\sum_{i=1}^{k}\sum_{j=0}^{m} w_j x_i^{n+j} = \sum_{i=1}^{k} y_i x_i^n \tag{41}$$

$$\begin{bmatrix} k & \sum_{i=0}^{k} x_i & \sum_{i=0}^{k} x_i^2 & \cdots & \sum_{i=0}^{k} x_i^n \\ \sum_{i=0}^{k} x_i & \sum_{i=0}^{k} x_i^2 & \sum_{i=0}^{k} x_i^3 & \cdots & \sum_{i=0}^{k} x_i^{n+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=0}^{k} x_i^m & \sum_{i=0}^{k} x_i^{m+1} & \sum_{i=0}^{k} x_i^{m+2} & \cdots & \sum_{i=0}^{k} x_i^{m+n} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{k} y_i \\ \sum_{i=0}^{k} y_i x_i \\ \vdots \\ \sum_{i=0}^{k} y_i x_i^m \end{bmatrix} \tag{42}$$

Or, in its **Closed Form**:

$$w^* = (X^T X)^{-1} X^T y \tag{43}$$

If we're using a set of $N \times D$ data, a nice and easy way to write the error function could be:

$$E_D(w) = \frac{1}{2}\sum_{n=1}^{N}\big\{t_n - w^T\phi(x_n)\big\}^2 \tag{44}$$

Where $\phi(x)$ represents the Gaussian density function. Remember that the above estimates are using **values** given to us, not **probabilities**. The relationship between the two is shown below:

$$p(t|x,w,\beta) = \prod_{n=1}^{N} \mathcal{N}\Big(t_n|y(x_n,w),\beta^{-1}\Big) = \prod_{n=1}^{N} \Big(\frac{\beta}{2\pi}\Big)^{\frac{1}{2}}\ exp\left\{\frac{-\beta\big(t_n - w^T\phi(x_n)\big)^2}{2}\right\} \tag{45}$$

$$\ln\, p(t|x,w,\beta) = \frac{N}{2}\,\ln\,\beta - \frac{N}{2}\,\ln(2\pi) - \frac{\beta}{2}\sum_{n=1}^{N}\big\{t_n - w^T\phi(x_n)\big\}^2 \tag{46}$$

$$w_{_{ML}} = (\Phi^T\Phi)^{-1}\Phi^T t_n \tag{47}$$

$$\mu_{_{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \text{ in a Gaussian PDF} \tag{48}$$

$$\Sigma_{_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\big(x_n - \mu_{_{ML}}\big)\big(x_n - \mu_{_{ML}}\big)^T \tag{49}$$

Remember that Φ is the CDF equivalent of ϕ, that is, $\Phi = \int \phi$.

2.5 Linear Models for Linear Regression

2.5.1 . . . know error function, basis function and design matrix.

See previous page.

2.5.2 . . . understand how regularization works.

The goal in regularization is to minimize:

$$E(w) = E_d(w) + \lambda E_w(w) \text{ where } \lambda \text{ is the regularization coefficient} \tag{50}$$

A simple form of a regularizer is:

$$E_W(w) = \frac{1}{2}w^T w \tag{51}$$

So the total error function becomes:

$$E(w) = \sum_{n=1}^{N}\Big\{t_n - w^T\phi(x_n)\Big\}^2 + \frac{\lambda}{2}w^T w \tag{52}$$

This is also called the **quadratic regularizer**.

2.5.3 . . . model complexity vs. performance

In polynomial curve fitting, an optimal order of polynomial gives the best generalization. The number of free parameters in the model, and therefore model complexity, is controlled by the order of the polynomial. If using regularized least squares, I also controls model complexity.

2.6 Linear Models for Classification

2.6.1 . . . understand least square method and know under what conditions it could fail

When working with classes, we assign input vector x to one of K classes, denoted by C_k. The two-class linear discriminant function is:

$$y(x) = w^t x + w_0 \tag{53}$$

We assign x to C_k if $y_k(x) \geq y_j(x)\ \forall\ j \neq k$. w is a weight vector, and w_0 is **bias**. So seach $C_k, k = 1, \ldots, K$ is described by it's own linear model $y_k(x) = w_k^T x + w_{k0}$. We create an augmented vector $x = (1, x^T)$ and $w_k = (w_{k0}, w_k^T)$, and simplify the notation to $y_k(x) = W^T x$. We then use the input values as the input vector $\vec{x}$, and assign x to the class for which the output is the largest. We determine W by minimizing squared error.

This is severely limited, as it is *very* sensitive to outliers.

2.6.2 . . . know what generative models and discriminative models are, and the difference between them

Simply put, a **Generative Model** learns the **joint** probability distribution $p(x, y)$, whereas a **discriminative** model learns the **conditional** probability distribution $p(y|x)$.

2.6.3 . . . know what the perceptron algorithm is and how to do the simple calculation.

The **perceptron algorithm** is simple: given an input vector x transformed by a fixed nonlinear transformation to give feature vector $\phi(x)$, then we have:

$$y(x) = f\big(w^T \phi(x)\big) \tag{54}$$

$$f(a) = \begin{cases} +1 & \text{if } a \geq 0 \\ -1 & \text{if } a < 0 \end{cases} \tag{55}$$

3 Lecture Notes

3.1 Machine Learning Overview

Machine Learning is a term given to programming computer to perform tasks that humans perform well, but are difficult to specify algorithmically. An example is recognizing handwritten digits. Handcrafted rules will result in a large number of rules and exceptions, so teaching a computer to recognize them is the easier and more efficient approach.

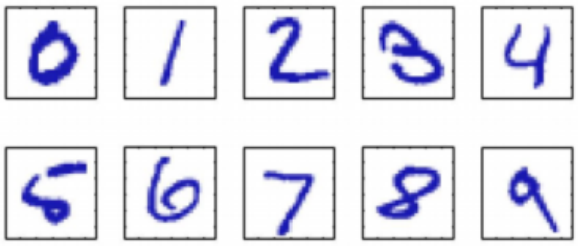


Figure 1: Wide variety of the same numeral

Machine learning takes a multi-tiered approach to solving problems.

- **Generalization**
  - **Data Collection** (Samples)
  - **Model Selection** (Probability distribution to model the process)
  - **Parameter Estimation** (Values / Distributions)
- **Decision**
  - **Inference** (Find responses to queries)

Popular Statistical Models include:

- **Generative**
  - Naïve Bayes
  - Mixtures of multinomials
  - Mixtures of Gaussians
  - Hidden Markov Models (HMM)
  - Bayesian Networks
  - Markov Random Fields
- **Discriminative**
  - Logistic Regression
  - SVMs
  - Traditional neural networks
  - Nearest neighbor
  - Conditional Random Fields (CRF)

3.2 Probability Theory

Probability is a the key concept when dealing with **uncertainty**. **Probability Theory** is the framework for quantification and manipulation of uncertainty.

3.2.1 An example

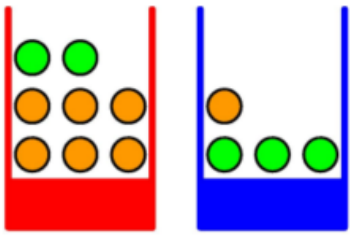


Figure 2: Graphical representation of the fruit-picking problem.

An example of a problem that deals with probability is picking a fruit out of a box while blind-folded. Which fruit is chosen is the random variable $F = \{o, a\}$, where o represents an orange, and a represents an apple. When looking at just the red box, we can say *the probability of picking an orange*, or $P(f = 0)$, is $\frac{3}{4}$, and *the probability of picking an apple*, or $P(f = a)$, is $\frac{1}{4}$.

The **probability distribution** is a mathematical function that describes:

1. The possible values of a random variable, and
2. the associated probabilities.

When dealing with Machine Learning, there are often several variables involved in a probability. In the fruit-picking example, we have two different probabilities – which box is chosen, or random variable $B = \{r, b\}$, and which fruit is chosen, or random variable $F = \{o, a\}$. In this example, we are choosing F to be dependent on B, that is, the probability of picking either fruit depends on which box was chosen. We say $P(B = r)$ is $\frac{4}{10} = \frac{2}{5}$, and $P(B = b)$ is $\frac{6}{10} = \frac{3}{5}$, as we are only choosing a *fruit* without knowledge of which box it is in.

When given this information, we can describe numerous probabilities of interest.

- **Marginal Probability** – What is the probability of picking an apple? $P(F = a)$
- **Conditional Probability** – We picked an orange. What is the probability that we chose it from the blue box? $P(B = b|F = o)$
- **Joint Probability** – What is the probability of picking an orange from the blue box? $P(B = b, F = 0)$

3.2.2 Sum Rule

The **Sum Rule of Probability Theory** states that, when considering two random variables X which can take on M different values $x_1, \ldots, X_M$ and Y which can take on L different values $y_1, \ldots, y_L$, and N trials sampling both X and Y, then taking $n_{(i,j)}$ which represents the number of trials where $X = x_i$ and $Y = y_j$, then the joint probability is:

$$p(X = x_i, Y = y_j) = \frac{n_{(i,j)}}{N} \tag{56}$$

Similarly, the marginal probability is:

$$p(X = x_i) = \frac{c_i}{N}, \text{ where } c_i = \sum_j n_{(i,j)}, \text{ or } P(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j) \tag{57}$$

3.2.3 Product Rule

The **Product Rule of Probability Theory**, instead, considers only those instances for which $X = x_i$. The fraction of instances where $Y = y_j$ as well is written as $P(Y = y_j|X = x_i)$ (read as *the probability $Y = y_j$ given that $X = x_i$*) and is called the **conditional probability**. This describes the relationship between joint and conditional probability as:

$$p(Y = y_j|X = x_i) = \frac{n_{(i,j)}}{c_i} \tag{58}$$

Combining the two equations we can see:

$$P(X = x_i, Y = y_j) = \frac{n_{(i,j)}}{N} = \frac{n_{(i,j)}}{c_i} \cdot \frac{c_i}{N} = p(Y = y_j|X = x_i) \cdot P(X = x_i) \tag{59}$$

3.2.4 Bayes' Rule

From the previous rules, together with the symmetry property $P(X, Y) = P(Y, X)$, we get **Bayes' Theorem**:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \text{ where } P(X) = \sum_Y P(X|Y)P(Y) \tag{60}$$

With this in hand, we can apply it to the fruit problem.

- Probability that we chose from the red box, given that we picked an orange:

$$\begin{aligned} P(B = r|F = o) &= \frac{P(F = o|B = r)P(B = r)}{P(F = o)} \\ &= \frac{\frac{3}{4} \times \frac{4}{10}}{\frac{9}{20}} = \frac{2}{3} = 0.66 \end{aligned}$$

- Probability that the fruit is an orange:

$$\begin{aligned} P(F = o) &= P(F = o, B = r) + P(F = o, B = b) \\ &= P(F = o|B = r)P(B = r) + P(F = o|B = b)P(B = b) \\ &= \frac{6}{8} \times \frac{4}{10} + \frac{1}{4} \times \frac{6}{10} = \frac{9}{20} = 0.45 \end{aligned}$$

3.2.5 Probability Densities

When x is a **Continuous Variable**, we can then use **Probability Densities**. If we say the probability that x falls in the interval $(x, x + \delta_x)$ is given by $P(X)dx$ for $\delta_x \to 0$, then $P(X)$ is a **Probability Density Function** of x. Then, the probability that x lies in the interval (a, b) is:

$$P(x \in (a, b)) = \int_a^b P(x)dx \tag{61}$$

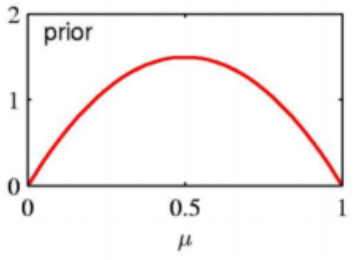


Figure 3: Probability that x lies in the interval $(-\infty, z)$ is $P(z) = \int_{-\infty}^{z} P(x)dx$

If there are several continuous variables $x_1, \ldots, x_D$ denoted by a vector $\vec{x}$, then we can define a **joint probability density** $P(x) = P(x_1, \ldots, x_D)$. It is important to note that a multivariate probability density must satisfy:

$$P(x) \leq 0 \text{ and } \int_{-\infty}^{\infty} P(x)dx = 1 \tag{62}$$

3.2.6 Expectation

Expectation is said to be the **average value** of some function $f(x)$ under the probability distribution $P(x)$, denoted as $E[f]$. For a discrete distribution:

$$E[f] = \sum_x P(x)f(x) \tag{63}$$

For a continuous distribution:

$$E[f] = \int P(x)f(x)dx \tag{64}$$

If there are N points drawn from a probability density function, then the expectation can be approximated as:

$$E[f] - \frac{1}{N}\sum_{n=1}^{N} f(x_n) \tag{65}$$

The **Conditional Expectation** with respect to a conditional distribution can be expressed as:

$$E_x[f] = \sum_x P(x|y)f(x) \tag{66}$$

3.2.7 Variance

Variance measures how much variability there is in $f(x)$ around its mean (or *expected*) value $E[f(x)]$. The variance of $f(x)$ can be calculated as:

$$var[f] = E[(f(x) - E[f(x)])^2] = E[f(x)^2] - E[f(x)]^2 \tag{67}$$

Similarly, the variance of the variable x itself is:

$$var[x] = E[x^2] - E[x]^2 \tag{68}$$

3.2.8 Covariance

For two random variables x and y, their **covariance** is defined as:

$$\begin{aligned} cov[x, y] &= E_{(x,y)}[\{x - E[x]\}\{y - E[y]\}] \\ &= E_{(x,y)}[xy] - E[x]E[y] \end{aligned}$$

This expresses how x and y vary together. If x and y are independent, then their covariance vanishes — similarly, if $\vec{x}$ and $\vec{y}$ are vectors of random variable, then their covariance is a matrix.

3.2.9 Bayesian Probabilities

The **classical** or **frequentist** view of probabilities is that probability can be described as *a frequency of random, repeatable events*. The **Bayesian** view, however, describes probability as a *quantification of uncertainty* — a degree of belief in propositions that do not involve random variables.

The use of probability to represent uncertainty is chosen out of necessity, rather than convenience. If numerical values are used to present degrees of belief, then a simple set of axioms for manipulating degrees of belief leads to the sum and product rules of probability described earlier. Probability can be regarded as *an extension of Boolean logic* to situations involving uncertainty.

The **Bayesian Approach** makes several qualifications to standard Probability Theory. The first distinction is that we quantify uncertainty around the choice of a parameter w, and uncertainty *before* observing data is expressed by $p(w)$. Given observed data $D = \{t_1, \ldots, t_N\}$, then uncertainty in w after observing D is given by Bayes' Rule as:

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

The quantity $p(D|w)$ can be viewed as a function of w, and represents *how probable the data set is for different parameters* w. This is called the **likelihood function**, and is **not** a probability distribution over w.

3.2.10 Likelihood Function

Bayes' rule, in words, could be described as:

$$posterior = \frac{likelihood \times prior}{evidence} \tag{69}$$

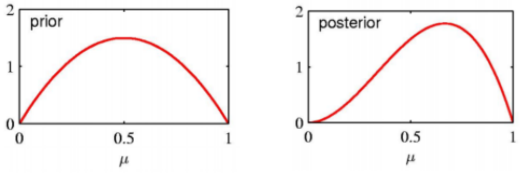
The likelihood function plays a central role in both Bayesian and frequentist paradigms. A frequentist might say that w is a fixed parameter determined by an estimators, with error bars ob estimate from possible data sets D. A Bayesian approach might say that there is a single data set D, and uncertainty is expressed as a probability distribution over w.

3.2.11 Maximum Likelihood Approach

In a frequentist setting, w is considered to be a fixed parameter, set to maximize the likelihood function $p(D|w)$. In machine learning, we describe the negative log of the likelihood function

$-log(p(D|w))$ as the **error function**, as maximizing likelihood is equivalent to minimizing error.

In the Bayesian approach, inclusion of prior knowledge arises naturally. As an example, say we take a coin and toss it three times, with it landing on heads each time. A classical approach with no prior knowledge would assume $P(heads) = 1$, implying all future coin tosses will land heads. A Bayesian approach will lead to a less extreme conclusion.



(a) Bayesian Approach (b) Classical Approach

Figure 4: Classical vs. Bayesian

3.2.12 Bayesian Approach

In choosing to approach a problem in the Bayesian fashion, it's important to note the factors that make it difficult. Marginalization over the *entire* parameter space is required to make predictions or compare models. However, sampling methods such as **Markov Chain** and **Monte Carlo** methods, as well as increased speed and memory of computers, help to offset these difficulties. As an alternative to sampling, deterministic approximation schemes such as **Variational Bayes** and **Expectation Propagation** can be used.

3.2.13 The Gaussian Distribution

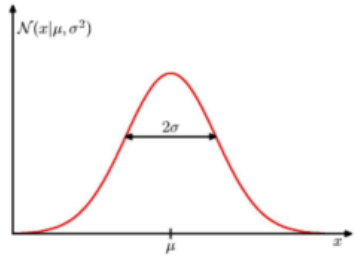


Figure 5: A Gaussian Distribution

Also called a **Normal Distribution**, this is a very common continuous probability distribution. For a single real-value variable x:

$$N(X|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{70}$$

Where $\mu = E[x]$ represents the **mean**, and $\sigma^2 = Var[x]$ the **variance**. Derived from this, $\sigma = \sqrt{\sigma^2}$ is the **standard deviation**, and $\beta = \frac{1}{\sigma^2}$ is the **precision**.

Given N observations $x_i, i = 1, \ldots, n$, independent and identically distributed, the probability of seeing these observations is given by the **likelihood function**:

$$p(x|\mu,\sigma^2) = \prod_{n=1}^{N} N(x_n|\mu,\sigma^2) \tag{71}$$

Similarly, the **log-likelihood function** is given by the formula:

$$ln\ p(x|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}ln\ \sigma^2 - \frac{N}{2}ln\ (2\pi) \tag{72}$$

Where the **maximum likelihood** equations are given by:

$$\mu_{_{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

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

Maximum likelihood systematically underestimates variance. $E[\mu_{_{ML}}] = \mu$, but $E[\sigma^2_{_{ML}}] = ((N-1)/N)\sigma^2$. This is because the variance is estimated relative to the sample mean, not the *true* mean. This is related to the **over-fitting problem.**

3.2.14 Probabilistic Curve Fitting

The goal of curve-fitting is to predict a target variable t given a new value of the input variable x. Given N input values $x = (x_1, \ldots, x_N)^T$ and corresponding targets value $t = (t_1, \ldots, t_N)^T$, we can assume that given a value of x, value t has a Gaussian distribution with a mean equal to $y(x, w)$ of the polynomial curve $p(t|x, w, \beta) = N(t|y(x, w), \beta^{-1})$, where the **mean** is given by the polynomial function $y(x, w)$ and the **precision** by β.

In regards to curve-fitting with maximum likelihood, we know the likelihood function is:

$$p(t|x, w, \beta) = \prod_{n=1}^{N} N(t_n|y(x_n, w), \beta^{-1}) \tag{73}$$

Similarly, the logarithm of the likelihood function is:

$$\ln\, p(t|x, w, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 + \frac{N}{2}\,\ln\,\beta - \frac{N}{2}\,\ln\,2\pi \tag{74}$$

To find the maximum likelihood solution for polynomial coefficients $w_{_{ML}}$, we first *maximize* with respect to w. We can omit the last two terms as they don't relate to w, and we can replace $\frac{\beta}{2}$ with $\frac{1}{2}$ for the same reason, and then minimize negative log-likelihood. This is identical to the sum-of-squares error function.

We can also use this to determine β of a Gaussian conditional distribution. Maximizing likelihood with respect to β gives us:

$$\frac{1}{\beta_{_{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, w_{_{ML}}) - t_n\}^2 \tag{75}$$

Note that we first need to determine the parameter vector $\vec{w}_{_{ML}}$ governing the mean. If we know the parameters w and β, predictions for new values of x can be made using:

$$p(t|x, w_{_{ML}}, \beta_{_{ML}}) = N(t|y(x, w_{_{ML}}, \beta^{-1}_{_{ML}}) \tag{76}$$

Instead of using a point estimate, we are now giving a probability distribution over t.

3.2.15 Posterior Distribution

Introducing *a prior* distribution over polynomial coefficients w gives us:

$$p(w|\alpha) = N(w|0, \alpha^{-1}I) = \left(\frac{a}{2\pi}\right)^{\frac{M+1}{2}} e^{-\frac{\alpha}{2}w^T w} \tag{77}$$

Where α is the precision of the distribution, and $M+1$ is the total number of parameters for an M^{th} degree polynomial. Using **Bayes' Theorem**, *posterior* distribution for w is proportional to the product of *prior* distribution and the likelihood function, or:

$$p(w|x, t, \alpha, \beta) = p(t|x, w, \beta)p(w|\alpha) \tag{78}$$

In this instance, w can be determined by finding the most probably value, *i.e.*, maximizing the posterior distribution. This is equivalent to minimizing:

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n.w) - t_n\}^2 + \frac{\alpha}{2}w^T w \tag{79}$$

This is the same as the sum of squared errors function with a regularization parameter given by $\lambda = \frac{\alpha}{\beta}$.

3.2.16 Bayesian Curve Fitting

Given training data x and t, and a new test point x, our goal is to predict the values of t, *i.e.*, we wish to evaluate the predictive distribution $p(t|x, x, t)$. If we apply the product and sum rules, we can see:

$$\begin{aligned} p(t|x,x,t) &= \int p(t,w|x,x,t)dw && \text{by Sum Rule} \\ &= \int p(t|x,w,x,t)p(w|x,x,t) && \text{by Product Rule} \\ &= \int p(t|x,w)p(w|x,t)dw && \text{by eliminating unnecessary variables} \end{aligned}$$

Where $p(t|x, w) = N(t|y(x, w), \beta^{-1})$, and $p(w|x, t)$ is the posterior distribution over parameters (Gaussian).

3.3 Model Selection

In polynomial curve fitting, an optimal order of polynomial gives the best generalization. The number of free parameters in the model, and therefore model complexity, is controlled by the order of the polynomial. If using regularized least squares, I also controls model complexity.

3.3.1 Validation Set to Select Model

Performance on a training set is not a good indicator of predictive performance. If there is plenty of data, some of the data can be used to train a range of models, or one given model is given a range of parameters. We can then compare this model with an independent set, called a **validation set**. The one with the best predictive performance will be the model chosen to represent the probability.

If a data set is small, then some over-fitting can occur.

3.3.2 S-fold Cross Validation

If a supply of data is limited, we can partition all available data into S groups. Of these groups, $S-1$ groups are used to train and then we evaluate with the remaining group. We repeat for all S choices of the validation group, and performance scores from S runs are averaged.

Figure 6: The red group is left out.

3.3.3 Bayesian Information Criterion

The **Bayesian Criterion** helps us choose a model. The **Akaike Information Criterion (AIC)** chooses the model for which the quantity $\ln p(D|w_{ML} - M$ is highest, where M is the number of adjustable parameters. The **BIC** is a variation of the quantity.

3.4 Polynomial Curve Fitting

3.4.1 Simple Regression Problem

We are given N observations of x, where $x = (x_1, \ldots, x_N)^T$ and $t = (t_1, \ldots, t_N)^T$. The goal is to exploit training set to predict a value of $\hat{t}$ from x — inherently this is a difficult problem, but probability theory allows us to make a prediction.

A polynomial problem is of the form:

$$y(x, w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{i=0}^{M} x_i x^i \tag{80}$$

In this form, M is considered the order of the polynomial. Note that a higher M value does not always yield a better fitting curve. Coefficients $w_0, \ldots, w_M$ are denoted by the vector $\vec{w}$. This is a **nonlinear** function of x, but a **linear** function of coefficients w.

3.4.2 Error Function

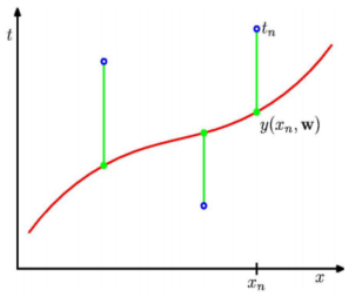


Figure 7: Red represents the best polynomial fit.

The **error function** is the sum of squares of the errors between the predictions $y(x_n, w)$ for each data point x_n and target value t_n. The formula for this is:

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2 \tag{81}$$

The factor of $\frac{1}{2}$ is included for later convenience. We solve for this equation by choosing the value of w for which $E(w)$ is as small as possible.

It's important to note that the error function is a **quadratic** in coefficients w, which means the derivative with respect to the coefficients will be linear in elements of w, thus, the error function has

a closed form solution. The unique minimum is denoted as w^*, and the resulting polynomial is $y(x, w^*)$. A solution takes the form:

$$\begin{aligned}\frac{\partial E(w)}{\partial w} &= \sum_{n=1}^{N}\{y(x_n, w) - t_n\}x_n^i \\ &= \sum_{n=1}^{N}\Big\{\sum_{j=0}^{M} w_j x_n^j - t_n\Big\}x_n^i\end{aligned}$$

After setting equal to zero ...

$$\sum_{n=1}^{N}\sum_{j=0}^{M} = \sum_{n=1}^{N} t_n x_n^i \tag{82}$$

Since

$$y(x, w) = \sum_{j=0}^{M} w_j x^j \tag{83}$$

3.4.3 Solving Simultaneous Equations

If an equation is of the form $aw = b$, where A is and $N \times (M+1)$ matrix, $\vec{w}$ is an $(M+1) \times 1$ vector, and $\vec{b}$ is an $N \times 1$ vector, then we can solve it using the matrix inversion $w = A^{-1}b$, or by **Gaussian elimination**.

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array} = x_1\begin{bmatrix} a_{11} \\ a_{12} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2\begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + x_n\begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \tag{84}$$

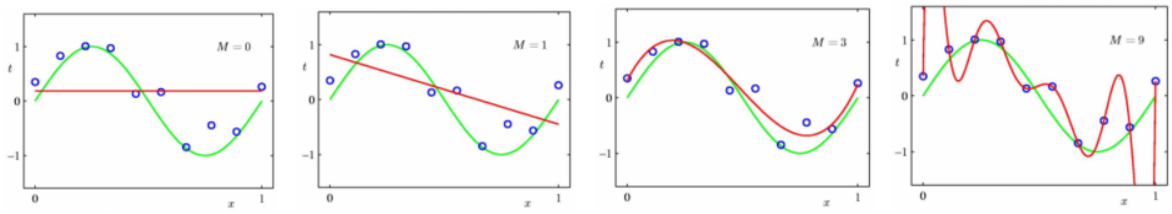
To break the equation down:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{12} & a_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \tag{85}$$

Following, let's see an example using Gaussian Elimination, followed by back-substitution:

$$\begin{aligned} x + 3y - 2z &= 5 \\ 3x + 5y + 6z &= 7 \\ 2x + 4y + 37 &= 8 \end{aligned}$$

$$\left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 3 & 5 & 6 & 7 \\ 2 & 4 & 3 & 8 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 0 & -4 & 12 & -8 \\ 2 & 4 & 3 & 8 \end{array}\right] \sim \left[\begin{array}{ccc|c} 1 & 3 & -2 & 5 \\ 0 & -4 & 12 & -8 \\ 0 & -2 & 7 & -2 \end{array}\right] \sim \cdots \sim \left[\begin{array}{ccc|c} 1 & 0 & 0 & -15 \\ 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & 2 \end{array}\right]$$



(a) $M = 0$ (b) $M = 1$ (c) $M = 3$ (d) $M = 9$

Figure 8: Choosing the order of M

	$M = 0$	$M = 1$	$M = 6$	$M = 9$
w_0^*	0.19	0.82	0.31	0.35
w_1^*		-1.27	7.99	232.37
w_2^*			-25.43	-5321.83
w_3^*			17.37	48568.31
w_4^*				-231639.30
w_5^*				640042.26
w_6^*				-1061800.52
w_7^*				1042400.18
w_8^*				-557682.99
w_9^*				125201.43

Figure 9: As M increases, magnitude of coefficients increases

3.4.4 Generalization Performance

Let's consider a separate test set of 100 points. Then, for each values of M, we evaluate:

$$E(w^*) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w^*) - t_n\}^2 \text{ where } y(x, w^*) = \sum_{j=0}^{M} w_j^* x^j \tag{86}$$

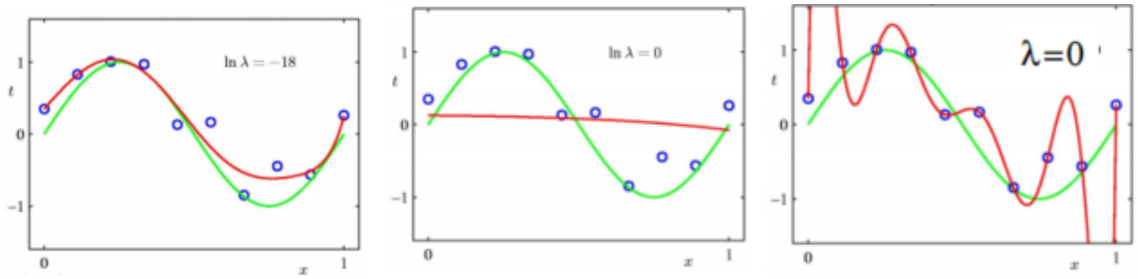
To evaluate the error, we want to use the **Root-Mean-Square** error:

$$E_{RMS} = \sqrt{\frac{2E(w^*)}{N}} \tag{87}$$

Division by N in the RMS error allows different sizes of N to be compared on equal footing, while the square root ensures E_{RM} is measured in the same units as t.

As we're developing models with higher M values, the coefficient w^* can change wildly. As an example:

For a given model, the problem of complexity overfitting becomes less severe as the size of the data set increases. A good rough rule would be that a data set should be at least 5 to 10 times as large as the number of parameters in the model.



(a) Optimal (b) Large Regularizer (c) No Regularizer

Figure 10: With $M = 9$, the effects of λ can be profound.

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

Figure 11: With $\ln \lambda = -\infty$, there is no regularizer. With $\ln \lambda = 0$, then $\lambda = 1$, which is a large regularizer.

3.4.5 Least Squares

If we limit the number of parameters to the size of the training set to avoid overfitting, we may be underestimating the complexity of the problem. In a real-world case with a much wider array of data, an overly simple model may be an inaccurate one. **Least Squares** is a specific case of Maximum Likelihood. The formula for least squares is:

$$\tilde{E}(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 + \frac{\lambda}{2}\ ||w^2|| \tag{88}$$

$$||w^2|| \equiv W^T w = w_o^2 + w_1^2 + \cdots + w_M^2 \tag{89}$$

This adds a *penalty term* to the error function to discourage coefficients from reaching large values – allowing a data set to be of limited size, but the model maintains necessary complexity. λ determines the relative importance of the regularization term to the error term. This method is also called **shrinkage** in statistics, or **weight decay** in neural networks.

The effect of the regularizer can be profound. See **Figure 10** for examples.

In these instances, λ controls the complexity of the model — hence it is an analogous choice to M. A suggested approach is to use a training set to determine coefficients for $\vec{w}$ using different values of M or λ, then use a validation set to optimize model complexity.

3.5 Discrete Probability Distributions

3.5.1 Bernoulli Distribution

A **Bernoulli Distribution** expresses a single binary-valued random variable, *e.g.*, $x \in \{0, 1\}$. The probability of $x = 1$ is denoted by a parameter μ, i.e.

$$p(x = 1|\mu) = \mu \tag{90}$$
$$p(x = 0|\mu) = 1 - \mu \tag{91}$$

The probability distribution has the form $Bern(x|\mu) = \mu^x(1 - \mu)^{1-x}$. The mean is shown to be $E[x] = \mu$, and the variance as $Var[x] = \mu(1 - \mu)$. The likelihood of n observations independently drawn from $p(x|\mu)$ is:

$$p(D|\mu) = \prod_{n=1}^{N} p(x_,|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1 - \mu)^{1-x_n} \tag{92}$$

Similarly, the log-likelihood is:

$$\ln\, p(D|\mu) = \sum_{n-1}^{N} \ln\, p(x_n|\mu) = \sum_{n=1}^{N}\{x_n \ln\, \mu + (1 - x_n) \ln\, (1 - \mu)\} \tag{93}$$

The maximum likelihood estimator is obtained by setting the derivative of $\ln\, p(D|\mu)$ with respect to m equal to zero, and is characterized by the equation:

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{94}$$

So, if the number of observations of $x = 1$ is m, then $\mu_{ML} = \frac{m}{N}$.

3.5.2 Binomial Distribution

The **Binomial Distribution** is related to the Bernoulli distribution. It expresses the distribution of m, and is proportional to $Bern(x|\mu)$. It adds up all the ways of obtaining the observation of $x = 1$, so:

$$Bin(m|N, \mu) = \binom{n}{m}\mu^m(1 - \mu)^{N-m} \tag{95}$$

The mean and variance are then:

$$E[m] = \sum_{m=0}^{N} m \cdot Bin(m|N, \mu) = N\mu \tag{96}$$
$$Var[m] = N\mu(1 - \mu) \tag{97}$$

As a reminder, for binomial coefficients,

$$\binom{N}{m} - \frac{N!}{m!(N - m)!} \tag{98}$$

3.5.3 Beta Distribution

Beta Distributions make use of the Gamma function.

$$Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-mu)^{b-1} \tag{99}$$

$$\Gamma(x) = \int_0^\infty \mu^{x-1}e^{-\mu}d\mu \tag{100}$$

Where a and b are hyperparameters that control the distribution of μ. Mean and variance are:

$$E[\mu] = \frac{a}{a+b} \tag{101}$$

$$var[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \tag{102}$$

The Maximum Likelihood Estimate in a Bernoulli distribution is the fraction of observations with $x = 1$, but is severely overfitted for small data sets. The likelihood function itself takes products of factors, in the form:

$$\mu^x(1-\mu)^{(1-x)} \tag{103}$$

If the prior distribution of μ is chosen to be proportional to power of μ and $1-\mu$, the posterior function will have the same functional form as the prior. This is called **Conjugacy**. The Beta function has a form suitable to a prior distribution $p(\mu)$. If we want a posterior function, we can multiply the Beta function with the binomial likelihood, yielding

$$p(\mu|m,l,a,b)\alpha\mu^{m+a-1}(1-\mu)^{l+b-1} = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1} \tag{104}$$

3.6 Multinomial Variables

3.6.1 Generalization of a Binomial

A **binomial** is something like tossing a coin — it expresses the probability of a number of successes in N trials. A **multinomial** is something like throwing a die — it expressed the probability of a given frequency for each value.

The Bernoulli distribution x is 0 or 1. If we want to generalize this to a variable that takes one of k values, we can use a schema where x is a **k-dimensional** vector. If $x = 3$, then $x = (0,0,1,0,0,0)^T$ is a possible representation with $k = 6$. Such vectors *must* satisfy $\sum_{i=1}^k x_k = 1$.

If the probability of $x_k = 1$ is denoted μ_k, then the distribution of x is given by the **Generalized Bernoulli**:

$$p(x|\mu) = \prod_{i=1}^k \mu_i^{x_i} \text{ where } \mu = (\mu_1,\ldots,\mu_k)^T \tag{105}$$

3.6.2 Maximum Likelihood Estimate of Generalized Bernoulli

If we have a data set D of N independent observations $x_1, \ldots, x_N$, where the n^{th} observation is written as $[x_{n1}, \ldots, x_{nk}]$, then the likelihood function has the form:

$$p(D|\mu) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{mk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{mn})} = \prod_{k=1}^{K} \mu_k^{m_k} \tag{106}$$

Where $m_k = \sum_n x_{nk}$ is the number of observations of $x_k = 1$. The maximum likelihood solution is obtained by setting the derivative with respect to μ, and is:

$$\mu_k^{ML} = \frac{m_k}{N} \tag{107}$$

3.6.3 Dirichlet Distribution

A **Dirichlet Distribution** is a family of prior distributions for parameters μ_k of a multinomial distribution. By inspection of the multinomial, the form of the conjugate prior is:

$$p(\mu|a)\alpha \prod_{k=1}^{K} \mu_k^{a_k - 1} \text{ where } 0 \leq \mu_k \leq 1 \text{ and } \sum_k \mu_k = 1 \tag{108}$$

Similarly, the normalized form of the Dirichlet Distribution is:

$$Dir(\mu|a) = \frac{\Gamma(a_0)}{\Gamma(a_1)\ldots\Gamma(a_k)} \prod_{k=1}^{K} \mu_k^{a_k - 1} \text{ where } a_0 = \sum_{k=1}^{K} a_k \tag{109}$$

3.6.4 Summary of Discrete Distributions

For **two-state** (binary) values, use a **Bernoulli** distribution. If there are two binary variables, use the **binomial**.

$$Bern(x|\mu) = \mu^x (1-\mu)^{1-x} \tag{110}$$

$$Bin(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \text{ where } \binom{N}{m} = \frac{N!}{m!(N-m)!} \tag{111}$$

For **k-states**, use the **Generalized Bernoulli** distribution. If there are many variables, use the **multinomial**.

$$p(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k} \text{ where } \mu = (\mu_1, \ldots, \mu_K)^T \tag{112}$$

$$Mult(m_1, m_2, \ldots, m_K|\mu, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k} \tag{113}$$

For **Conjugate Priors**, if it is binomial, use the **Beta** distribution, if multinomial, use **Dirichlet**.

$$Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(a-\mu)^{b-1} \tag{114}$$

$$Dir(\mu|a) = \frac{\Gamma(a_0)}{\Gamma(a_1)\Gamma(a_2)\cdots\Gamma(a_K)} \prod_{k=1}^{K} \mu_k^{a_k - 1} \text{ where } a_0 = \prod_{k=1}^{K} a_k \tag{115}$$

3.7 Deep Dive: The Gaussian Distribution

As we've stated before, for a single real-values variable x, with parameters μ (mean) and σ^2 (variance):

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{116}$$

As well, σ is the *standard deviation*, $\beta = \frac{1}{\sigma^2}$ is the precision, $E[x] = \mu$ and $Var[x] = \sigma^2$. For a D-dimensional vector $\vec{x}$, the multivariate Gaussian is:

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}} \times \frac{1}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{117}$$

Where μ is a mean vector, Σ is a $D \times D$ covariance matrix, and $|\Sigma|$ is the determinant of Σ. As well, Σ^{-1} is sometimes referred to as the **precision matrix**.

The **Covariance Matrix** is a measure of the dispersion of the data. The element in position i, j represents the covariance between the i^{th} and j^{th} variables, written as $E[(x_i - \mu_i)(y_i - \mu_j)]$.

3.7.1 Importance of the Gaussian

The Gaussian distribution arises in many contexts. For a given variable, the Gaussian maximizes entropy. If you take a set of random variable and sum them, they become increasingly Gaussian in nature.

3.7.2 Maximum Likelihood for the Gaussian

Given a data set $X = (x_1, \ldots, x_N)^T$ where the observations $\{x_n\}$ are drawn independently, the log-likelihood functions is given by:

$$\ln p(X|\mu,\Sigma) = -\frac{ND}{2} \ln (2\pi) - \frac{N}{2} \ln |\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^T\Sigma^{-1}(x_n - \mu) \tag{118}$$

The derivative with respect to μ:

$$\frac{\partial}{\partial\mu} \ln p(X|\mu,\Sigma) = \sum_{n=1}^{N}\Sigma^{-1}(x_n - \mu) \tag{119}$$

The solution to this equation is:

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{120}$$

Maximization with respect to Σ is slightly more involved. It yields:

$$\Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})(x_n - \mu_{ML}^T \tag{121}$$

4 Homework Problem Sets

4.1 Homework Set 1

4.1.1 Independence, Marginal and Conditional Probabilities

Consider the joint distribution of x and y described in the table below.

x/y	0	1
0	2/9	4/9
1	1/9	2/9

(a) Calculate the marginal probabilities given different values of x and y.

i. $p(x=0) = \sum_{n=1}^{N} p(x=0|y) = \frac{2}{9} + \frac{4}{9} = \frac{2}{3}$

ii. $p(x=1) = \sum_{n=1}^{N} p(x=1|y) = \frac{1}{9} + \frac{2}{9} = \frac{1}{3}$

iii. $p(y=0) = \sum_{n=1}^{N} p(y=0|x) = \frac{2}{9} + \frac{1}{9} = \frac{1}{3}$

iv. $p(y=1) = \sum_{n=1}^{N} p(y=1|x) = \frac{4}{9} + \frac{2}{9} = \frac{2}{3}$

(b) Calculate the conditional probabilities given different values of x and y.

i. $p(x=0|y=0) = \frac{p(x=0,y=0)}{p(y=0|x)} = \frac{2/9}{\sum_{n=1}^{N} p(y=0|x)} = \frac{2/9}{2/9+1/9} = \frac{2/9}{3/9} = \frac{2}{3}$

ii. $p(x=0|y=1) = \frac{p(x=0,y=1)}{p(y=1|x)} = \frac{4/9}{\sum_{n=1}^{N} p(y=1|x)} = \frac{4/9}{4/9+2/9} = \frac{4/9}{6/9} = \frac{2}{3}$

iii. $p(x=1|y=0) = \frac{p(x=1,y=0)}{p(y=0|x)} = \frac{1/9}{\sum_{n=1}^{N} p(y=0|x)} = \frac{1/9}{2/9+1/9} = \frac{1/9}{3/9} = \frac{1}{3}$

iv. $p(x=1|y=1) = \frac{p(x=1,y=1)}{p(y=1|x)} = \frac{2/9}{\sum_{n=1}^{N} p(y=1|x)} = \frac{2/9}{4/9+2/9} = \frac{2/9}{6/9} = \frac{1}{3}$

v. $p(y=0|x=0) = \frac{p(y=0,x=0)}{p(x=0|y)} = \frac{2/9}{\sum_{n=1}^{N} p(x=0|y)} = \frac{2/9}{4/9+2/9} = \frac{2/9}{6/9} = \frac{1}{3}$

(c) Random variables x and y __are__(are/aren't) independent.

4.1.2 The Gaussian Distribution and its Properties

For each of the questions below, choose one of the listed answers and fill in the blank.

(a) If $X \sim \mathcal{N}(x|0,1)$

i. $p(|X| < 1) =$ __A__

ii. $p(|X| < 2) =$ __B__

iii. $p(|X| < 3) =$ __C__

A) 0.683

B) 0.954

C) 0.997

Explanation: $X \sim \mathcal{N}(x|0,1)$ means X is approximately simulated by the **standard normal distribution**. As a reminder, the arguments of $\mathcal{N}$ are μ and σ, for *mean* and *standard deviation*. This means that the *expected value* for X is 0, so $p(|X| < 1)$, means,

"how many values fall within $(\mu - 1)$ and $(\mu + 1)$?" As the standard deviation is 1, this becomes, "how many values fall within $(\mu - \sigma)$ and $(\mu + \sigma)$, or "how many values fall within one standard deviation of the mean?" We know that 68% of values fall within one standard deviation of the mean, 95% within two, and 99.7% within three, which is what is described by A, B, and C, respectively.

(b) If $X \sim \mathcal{N}(x|0, 1)$, for any given a $(0 < a < 1)$, there exists U_a such that $p(X > U_a) = a$.

i. If $a = \frac{1}{2}$, then $U_a =$ A

A. 0
B. 1/4
C. 1/2
Explanation: If the probability that X is greater than some value is $1/2$, we know that value *must be the mean.* Since $\mu = 0$, the answer is A.

ii. $p(X < U_a) =$ B

A. $2a$
B. $1 - a$
C. a
Explanation: If $p(X > U_a) = a$, then $p(X < U_a) = 1 - P(X > U_a) = 1 - a$.

iii. If we know $p(X < x) = a$, then $x =$

A. U_{1-a}
B. $U_{1-a/2}$
C. $U_a/2$
Explanation: This is the same reason as above.

4.1.3 Bayes Rule

A die is selected at random from two 20-faced dice on which the symbols 1-10 are written with non-uniform frequency as follows:

Symbol	1	2	3	4	5	6	7	8	9	10
Number of faces of die A	6	3	3	2	2	1	1	1	1	0
Number of faces of die B	3	3	2	2	2	2	2	2	1	1

The randomly chosen die is rolled 8 times, with the following outcome:

$$D = \{5, 3, 9, 3, 8, 4, 7, 6\}$$

(a) Calculate the probability of getting the outcomes for each die.

i. $p(D|A) =$ 9/6.4e10

$$p(D|A) = p(5|A) \times p(3|A) \times p(9|A) \times p(3|A) \times p(8|A) \times p(4|A) \times p(7|A) \times p(6|A)$$
$$p(D|A) = \frac{2}{20} \times \frac{3}{20} \times \frac{1}{20} \times \frac{3}{20} \times \frac{1}{20} \times \frac{2}{20} \times \frac{1}{20} \times \frac{1}{20}$$
$$p(D|A) = \frac{2 \times 3 \times 3 \times 2}{20^8} = \frac{36}{25600000000} = \frac{9}{64000000000}$$

ii. $p(D|B) =$ 1/2e9

$$p(D|B) = p(5|B) \times p(3|B) \times p(9|B) \times p(3|B) \times p(8|B) \times p(4|B) \times p(7|B) \times p(6|B)$$
$$p(D|B) = \frac{2}{20} \times \frac{2}{20} \times \frac{1}{20} \times \frac{2}{20} \times \frac{2}{20} \times \frac{2}{20} \times \frac{2}{20} \times \frac{2}{20}$$
$$p(D|B) = \frac{2^7}{20^8} = \frac{1}{200000000}$$

(b) Given the outcome, calculate the probability that the die rolled was die A.

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)}$$
$$P(A|D) = \frac{P(D|A)P(A)}{P(D|A)P(A) + P(D|B)P(B)}$$
$$P(A|D) = \frac{\frac{36}{20^8} \times \frac{1}{2}}{\frac{36}{20^8} \times \frac{1}{2} + \frac{2^7}{20^8} \times \frac{1}{2}}$$
$$P(A|D) = \frac{18}{20^8} \times \frac{1}{\frac{18}{20^8} + \frac{64}{20^8}}$$
$$P(A|D) = \frac{18}{20^8} \times \frac{20^8}{82}$$
$$P(A|D) = \frac{18}{82} = \frac{9}{41}$$

4.2 Homework Set 2

4.2.1 Polynomial Curve Fitting

You are given a table with the data shown below.

k	x_k	y_k
1	1	4
2	2	4.5
3	3	6
4	4	8
5	5	8.5

(a) Fit the points using the function below. *(i.e., minimize E)*

$$E = \frac{1}{2} \sum_{i=0}^{k} (\hat{y}_k(w; x) - y_k)^2)$$

i. If we use the form of the linear function $\hat{y}(w; x) = w_0 + w_1 x$ then the fitting function is $\hat{y}(x) = \frac{5}{4}x + \frac{49}{20}$

Explanation: If we are using the form of $w_0 + w_1 x$, we are really using the familiar $y = ax + b$ linear formula. We know that:

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

So,

$$E = \frac{1}{2}\sum_{i=1}^{k}(\hat{y}_k(w;x) - y_k)^2)$$
$$= \frac{1}{2}\sum_{i=1}^{k}\{\sum_{j=0}^{1} w_j x_i^j - y_i\}^2$$

If we take the derivative:

$$\frac{\partial E(w)}{\partial w} = \sum_{i=1}^{k}\{\sum_{j=0}^{1} w_j x_j^i - y_k\}x_i^n$$

And set it equal to zero:

$$\sum_{i=1}^{k}\sum_{j=0}^{1} w_j x_i^{n+j} = \sum_{i=1}^{k} y_i x_i^n$$
$$\sum_{i=1}^{k} w_0 x_i + w_1 x_i^2 = \sum_{i=1}^{k} y_i x_i$$
$$\begin{bmatrix} k & \sum_{i=0}^{k} x_i \\ \sum_{i=0}^{k} x_i & \sum_{i=0}^{k} x_i^2 \end{bmatrix}\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{k} y_i \\ \sum_{i=0}^{k} x_i y_i \end{bmatrix}$$
$$\begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 31 \\ 105.5 \end{bmatrix}$$

After row reduction:

$$\left[\begin{array}{cc|c} 1 & 0 & 2.45 \\ 0 & 1 & 1.25 \end{array}\right]$$

ii. If we use the form $\hat{y}(w;x) = w_0 + w_1 x + w_2 x^2$, then the fitting function is $\hat{y}(x) = \underline{\frac{1}{28}x^2 + \frac{29}{28}x + \frac{27}{10}}$

$$\begin{bmatrix} k & \sum_{i=0}^{k} x_i & \sum_{i=0}^{k} x_i^2 \\ \sum_{i=0}^{k} x_i & \sum_{i=0}^{k} x_i^2 & \sum_{i=0}^{k} x_i^3 \\ \sum_{i=0}^{k} x_i^2 & \sum_{i=0}^{k} x_i^3 & \sum_{i=0}^{k} x_i^4 \end{bmatrix}\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{k} y_i \\ \sum_{i=0}^{k} x_i y_i \\ \sum_{i=0}^{k} x_i^2 y_i \end{bmatrix}$$
$$\begin{bmatrix} 5 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix}\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 31 \\ 105.5 \\ 416.5 \end{bmatrix}$$

After row reduction,

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 2.7 \\ 0 & 1 & 0 & 1.03571 \\ 0 & 0 & 1 & 0.0357143 \end{array}\right]$$

(b) If we are trying to fit the points using a polynomial of degree M of the form $\hat{y}(w;x) = w_0 + w_1 x^2 + \cdots + w_M x^M$, the error function takes the form $E = \frac{1}{2}\sum_{i=0}^{k}(\hat{y}_k(w;k) - y_k)^2)$. Derive a closed form solution for the parameters w defined to be $[w_0, w_1, \ldots, w_M]^T$. *Hint: represent your answer in Matrix form. Use* $X = (x_i^j)_{n\times m}$ *and* $y = [y_0, \ldots, y_n]^T$ *to express your*

answer. $w = \underline{(X^T X)^{-1} X^T y}$

Explanation: In our previous answers, we opted for using Gaussian Eliminations to find the answers, however, another option would be to make one side of the equation solely the $\vec{w}$ vector, and calculate the opposite as a multiplication of A^{-1} and y. What is important to remember is that we took the derivative of the original function in order for our solution to be *linear* in terms of $\vec{w}$. This means we removed a factor of X, which, if we are not taking the derivative, needs to be put back into the function. Thus X^T represents a column vector of x, and in putting this back in brings us X, a n $N \times M$ matrix, inverted, multiplied by y, both with a column vector X^T returned.

(c) We can modify the error function by introducing a regularization factor, *i.e.*, $E = \frac{1}{2}\sum_{i=1}^{k}(\hat{y}_k(w;x) - y_k)^2) + \frac{1}{2}|W|^2$. Fit the data above using the function and the information below:

i. If we use the linear form of the function, $\hat{y}(w;x) = w_0 + w_1 x$, then the fitting function is $\hat{y}(x) = \underline{\frac{307}{222} + \frac{56}{37}x}$

References

[1] Chrstopher M. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, 2006

[2] Sargur N. Srihari, *CSE474/574 Machine Learning Class Notes*. University at Buffalo, Fall 2015.