Chapter 1 Introduction

1.1 Example: Polynomial Curve Fitting

Example. *Given* $\mathbf{x} = (x_1, \cdots, x_N)^T$ *where* $x_n \in [0, 1]$ *and* $\mathbf{t} = (t_1, \cdots, t_N)^T$. *We generate target with*

$$t_n = \sin(2\pi x_n) + \epsilon$$

where ϵ *are Gaussian noises. The dataset capture the property that it possess an underlying regularity, which we wish to learn, but that individual observations are corrupted by random noise. We fit the data with a polynomial function of the form*

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

where M *is order of the polynomial. Functions which are linear in the unknown parameters are called* ***linear models****. Value of coefficients determined by minimizing an* ***error function*** *measuring the misfit between function* $y(x, \mathbf{w})$, *for a given* $\mathbf{w}$, *and the training set data points*

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

We find $\mathbf{w}^*$ *such that* $E(\mathbf{w}^*)$ *is minimized. Over-fitting problem can be understood as a general property of maximum likelihood. One etchnique used to contrl over-fitting phenomenon is* ***regularization****, which involves adding a penalty term to the error function in order to discourage the coefficients from reaching large values*

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

where $\|\mathbf{w}\| = \mathbf{w^T w} = w_0^2 + w_1^2 + \cdots + w_M^2$. *Techniques such as this are known as* ***shrinkage*** *methods because they reduce the value of coefficients. The particular case of a quadratic regularizer is called* ***ridge regression*** *or* ***weight decay*** *in the context of neural networks.* λ *controls the effective complexity of the model and hence determines the degree of over-fitting. We can divide data into* ***training set****, used to determine coefficients* $\mathbf{w}$, *and a separate* ***validation set****, used to optimize model complexity (either* M *or* λ*).*

1.2 Probability Theory

Definition. ***Sum and Product Rules of Probability*** *Let* $X \in \{x_i\}$ *where* $i = 1, \cdots, M$ *and* $Y \in \{y_j\}$, *where* $j = 1, \cdots, L$. *Consider a total of* N *trials in which we sample both* X *and* Y, *and let number of such trials in which* $X = x_i$ *and* $Y = y_j$ *be* n_{ij}. *Let number of trials in which* X *takes the value* x_i *be* c_i *and let number of trials in which* Y *takes* y_j *be denoted by* r_j. *So we have joint, marginal, and conditional probability as follows*

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

We have

$$p(X) = \sum_Y p(X, Y) \qquad \textbf{(sum rule)}$$

$$p(X, Y) = p(Y|X)p(X) \qquad \textbf{(product rule)}$$

Together we have ***Bayes' theorem***

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

where we have A as proposition and B as fixed evidence. $P(A|B)$ is called the ***posterior****, or the degree of belief after observing B. $P(A)$ is called* ***prior****, the initial degree of belief on A available before the observation. and ${}^{p(B|A)}/_{p(B)}$ is the* ***likelihood*** *supporting B given A. In essence, we can interpret the formula as posterior is proportional to prior times likelihood.*

Definition. ***Probability Density*** *Probability with respect to continuous variables. If probability of a real-valued variable x falling in interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta \to 0$, then $p(x)$ is the* ***probability density*** *over x.*

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

satisfying

$$p(x) \geq 0 \qquad \int_{-\infty}^{\infty} p(x)dx = 1$$

Cumulative distribution function *is the probability that x lies in the interval $(-\infty, z)$*

$$P(z) = \int_{-\infty}^{z} p(x)dx$$

which satisfies $P'(x) = p(x)$.
For several continuous variabes $x_1, \cdots, x_D$, denoted by $\mathbf{x}$, we define a ***joint probability density*** *as $p(\mathbf{x}) = p(x_1, \cdots, x_D)$ be probability of $\mathbf{x}$ falling in an infinitesimal volumne $\delta\mathbf{x}$ containing the point given by $p(\mathbf{x})\delta(x)$. The multivariate probability density satisfy*

$$p(\mathbf{x}) \geq 0 \qquad \int p(\mathbf{x})d\mathbf{x} = 1$$

The ***sum and product rule*** *for continous variables are given by*

$$p(x) = \int p(x, y)dy \qquad p(x, y) = (y|x)p(x)$$

Definition. ***Expectations and Covariances***

1. ***Expectation*** *The average value of some function $f(x)$ under a probability distribution $p(x)$ is called the expectation of $f(x)$, denoted by $\mathbb{E}\{f\}$*

$$\mathbb{E}\{f\} = \sum_x p(x)f(x) \qquad\qquad \mathbb{E}\{f\} = \int p(x)f(x)dx$$

2. ***Expectation Over Several Variables*** *We use subscript to indicate which variable is being averaged over so $\mathbb{E}\{f(x,y)\}$ denotes the average of the function $f(x,y)$ with respect to the distribution of x, but note it is a function of y.*

3. ***Conditioanl Expectation*** *The conditional expectation with respect to a conditional distribution is given by*

$$\mathbb{E}_x\{f|y\} = \sum_x p(x|y)f(x) \qquad\qquad \mathbb{E}_x\{f|y\} = \int p(x|y)f(x)dx$$

4. ***Variance*** *is a measure f how much variability there is in $f(x)$ around its mean value $\mathbb{E}\{f(x)\}$*

$$var\{f\} = \mathbb{E}\{(f(x) - \mathbb{E}\{f(x)\})^2\} = \mathbb{E}\{f(x)^2\} - \mathbb{E}\{x\}^2$$

5. ***Covariance*** *is a measure to which x and y vary together, covariance is zero if x and y are independent*

$$cov\{x,y\} = \mathbb{E}_{x,y}\{(x - \mathbb{E}\{x\})(y - \mathbb{E}\{y\})\} = \mathbb{E}_{x,y}\{xy\} - \mathbb{E}\{x\}\mathbb{E}\{y\}$$

For vectors $\mathbf{x}, \mathbf{y}$

$$cov\{\mathbf{x},\mathbf{y}\} = \mathbb{E}_{\mathbf{x},\mathbf{y}}\{(\mathbf{x} - \mathbb{E}\{\mathbf{x}\})(\mathbf{y^T} - \mathbb{E}\{\mathbf{y^T}\})\} = \mathbb{E}_{\mathbf{x},\mathbf{y}}\{\mathbf{xy^T}\} - \mathbb{E}\{\mathbf{x}\}\mathbb{E}\{\mathbf{y^T}\}$$

Note $cov\{x\} = cov\{x,x\}$

Definition. ***Bayesian Probabilities***

1. ***Motivation*** *For uncertain events which does not repeat numerous times (like arctic ice cap melt by end of centuary ...). Idea is we want to quantify expression of uncertainty and make precise revisions of uncertainty in light of new evidence, as well as subsequently be able to make optimal actions or decisions as a consequence. Then central idea is the use of probability to represent uncertainty.*

2. ***Bayes' Theorem*** *We capture assumptions about $\mathbf{w}$, before observing data in the form of **prior distribution** $p(\mathbf{w})$. The effect of observed data $\mathcal{D} = \{t_1, \cdots, t_N\}$ is expressed through $p(\mathcal{D}|\mathbf{w})$. We can evaluate the uncertainty in $\mathbf{w}$ after we have observed $\mathcal{D}$ in the form of **posterior probability** $p(\mathbf{w}|\mathcal{D})$*

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

*$p(\mathcal{D}|\mathbf{w})$ is evaluated for the dataset, and is a function of the parameter vector $\mathbf{w}$, in which case it is called the **likelihood function**. It expresses how probable that observed data is for different settings of parameter vector $\mathbf{w}$. $p(\mathcal{D})$ is a normalizing constaint ensuring that posterior distribution is a valid probability density (integrates to 1)*

$$posterior \propto likelihood \times prior$$

all of which are functions of $\mathbf{w}$. Bayesian and frequentist interpret likelihood $p(\mathcal{D}|\mathbf{w})$ differently

3. ***Frequentist*** $\mathbf{w}$ *is a fixed parameter, whose value is determined by some estimator, whose distribution is obtained by considering distribution of possible datasets $\mathcal{D}$. Maximum likelihood estimator chooses $\mathbf{w}$ for which the probability of observed dataset is maximized, i.e. maximizes the likelihood function $p(\mathcal{D}|\mathbf{w})$. The neagative logarithm of likelihood function is the **error function**.*

4. ***Bayesian*** *There is only a single dataset $\mathcal{D}$, and the uncertainty in the parameter is expressed through a probability distribution over $\mathbf{w}$. One advantage of Bayesian viewpoint is the inclusion of prior konwledge. Critics of Bayesian approach states that prior distribution is often selected based on mathematical convenience rather than as a reflection of prior beliefs. The Bayesian framework is limited by the difficulty in carrying out Bayesian procedure, particularly in the need to marginalize over the whole parameter space. The development of Markov chain Monte Carlo opens up to practical use of Bayesian techniques. Variational Bayes and Expectation progapation developed for deterministic approximation*

Definition. ***The Gaussian Distribution***

1. ***Single-Valued***

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

where μ is the mean, σ^2 is the variacne, $\beta = {}^1/_{\sigma^2}$ is called the precision. Note

$$\mathcal{N}\left(x|\mu,\sigma^2\right) > 0 \qquad \int_{-\infty}^{\infty} = \mathcal{N}\left(x|\mu,\sigma^2\right) dx = 1$$

with

$$\mathbb{E}\{x\} = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x dx = \mu \qquad \mathbb{E}\{x^2\} = \mu^2 + \sigma^2 \qquad var\{x\} = \sigma^2$$

2. ***Multivariate***

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

where D-dimensional vector $\boldsymbol{\mu}$ is called the mean and $D \times D$ matrix $\boldsymbol{\Sigma}$ is called the covaraince. $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$

3. ***Likelihood Function*** *Let* $\mathbf{x} = (x_1, \cdots, x_N)^T \overset{i.i.d.}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$

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

We maximize log likelihood,

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

We derive the maximum likelihood estimators

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

Note maximum likelihood approach systematically underestimates variance, because of introduction of biases. Note maximum likelihood variance underestimates the true variance

$$\mathbb{E}\{\mu_{mle}\} = \mu \qquad\qquad \mathbb{E}\{\sigma^2_{mle}\} = \frac{N-1}{N}\sigma^2$$

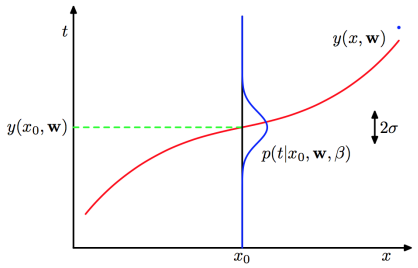
An unbiased estimator is given by

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

For large sample, this is not a problem. In models where there are many parameters, this problem is much more severe. Issue of bias is root of overfitting problem

4. ***Polynomial Fitting Revisited*** *We want to express uncertainty over value of target variable using a probability distribution. We assume, given* x*, the corresponding value of* t *has Gaussian distirbution with mean equal to* $y(x, \mathbf{w})$*, so*

$$p(\mathcal{D}|\mathbf{w}) = p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t \mid y(x, \mathbf{w}), \beta^{-1}\right)$$



Given training data $\{\mathbf{x}, \mathbf{t}\}$*, we want to determine values for* $\mathbf{w}$ *and* β *by maximum likelihood. The likelihood function is given by*

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

Maximizing likelihood is equivalent to minimizing sum-of-squares error function

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

We can find maximum likelihood estimator $\mathbf{w}_{mle}$ *and the corresponding precision given by*

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

Having determined parameter $\mathbf{w}$ *and* β*, we can make predictions for new values of* x *by expressing in terms of the predictive distribution.*

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

We can introduce a prior distribution over polynomial coefficients $\mathbf{w}$

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

where α *is precision of prior distribution and* $M+1$ *is the total number of elements in* $\mathbf{w}$ *for* M*th order polynomial. Also note* $\mathbf{\Sigma}^{-1} = (\alpha^{-1}\mathbf{I})^{-1} = \alpha\mathbf{I}$*. Using Bayes' theorem, we have*

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

We can determine $\mathbf{w}$ *by finding most probable value of* $\mathbf{w}$ *given the data, by maximizing the posterior distribution. Finding* ***maximum posterior*** *(MAP) is equivalent to minimizing negative logarithm of*

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

which is equivalent to minimize regularized sum-of-squares erorr fucntion with $\lambda = \alpha/\beta$

Definition. ***Model Selection***

1. ***Motivation*** *Order of polynomial or regularization coefficients govern complexity of model. We want to determine the value of hyperparameters such that we achieve best predictive performance of new data. Idea is to use an independent data from* ***training set****, called* ***validation set****, and select one having the best performance. However if model is iterated many times, then some overfitting to validation data can occur., so may want to keep a third* ***test set*** *on which performance of selected model is evaluated.*

2. ***Cross-Validation*** *Allows proportion of* $(S-1)/S$ *of data to be used for training. The drwaback is*

   (a) *Number of training runs increased by a factor of* S*, and can prove problematic for model in which the training itself is computationally expensive.*

   (b) *If there are multiple complexity parameter in the model, exploring combinations of settings of such parameter could, in worst case, require a number of training runs that is exponential in the number of parameters.*

   So want to find a measure of performance which depends only on the trainin gdata and which does not suffer from bias due to overfitting. AIC/BIC are such models

1.4 The Curse of Dimensionality

Definition. ***Curse of Dimensionality***

1. ***Grid Search*** *Divide region of space into regular cells, number of cell grow exponential with dimensionality of space. Need exponentially large quantity of training data to ensure cell not empty*

2. ***Polynomial Fitting*** *As dimension* D *increases, number of independent coefficients grows proportionally to* D^n *for* n*-degree polynomial*

1.5 Decision Theory

Definition. ***Motivation***

1. ***Statistical inference*** *is the process of deducing properties of an underlying probability distribution by analysis of data. More concretely, it deals with determination of* $p(\mathbf{x}, \mathbf{t})$*, the joint distribution providing a complete summary of uncertainty associated with these variables.*

2. ***Decision theory*** *deals with making optimal decisions given the appropriate probabilities, which is usually simple once we solved the inference problem.*

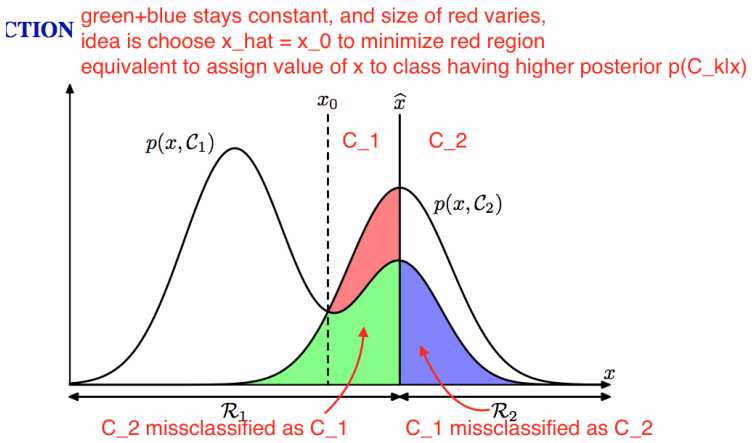
Definition. ***Minimizing Misclassification Rate*** *We want to come up with a rule that divide the input space into regiosn* $\mathcal{R}_k$ *called decision regions, one for each class, such that all points in* $\mathcal{R}_k$ *are assigned to class* $\mathcal{C}_k$*. The boundary between them are called decision boundaries. The probability of missclassification which we want to minimize is given by*

$$p(mistake) = p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) = \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x}$$

Generally for K *classes, it's easier maximize the probability of being correct, given by*

$$p(correct) = \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k) = \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} \propto \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathcal{C}_k | \mathbf{x}) d\mathbf{x}$$

Since $p(\mathbf{x})$ is common to all terms, we see that each $\mathbf{x}$ should be assfigned to class having largest posterior probability $p(C_k|\mathbf{x})$.



Definition. ***Minimizing the Expected Loss***

1. ***Loss Function*** *is a single, overall measure of loss incurred in taking any of the available decisions or actions.*

2. ***Loss Matrix*** *For a new value of $\mathbf{x}$, the true class is $\mathcal{C}_k$ and we assign $\mathbf{x}$ to class $\mathcal{C}_j$, incurring a loss that we denote by L_{kj}, an element in the loss matrix.*

3. ***Expected Loss*** *is the sum of the values of all possible losses, each multiplied by the probability of that loss occurring. Note loss function depends on the true class, which is unknown, so we instead try to minimize the expected loss*

$$\mathbb{E}\{L\} = \sum_k \sum_j \int_{x \in \mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x}$$

Note each $\mathbf{x}$ is independently assigned to one of decision region $\mathcal{R}_j$. The goal is to pick a region $\mathcal{R}_j$ in order to minimize the expected loss. The decision rule that minimizes the expected loss is one that assigns each new $\mathbf{x}$ to the class j for which

$$\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_j) \propto \sum_k L_{kj} p(\mathcal{C}_j|\mathbf{x})$$

is minimum.

Definition. ***The rejection option*** *Classification erorr arise form regions of input space wherew largest of posterior probability $p(\mathcal{C}_k|\mathbf{x})$ is significantly less than unity, or equivalently where the joint distribution $p(\mathbf{x}, \mathcal{C}_k)$ have comparable values. These are the region we are uncertain about class membership.*

Definition. ***Inference and Decision***
3 approaches to solve the decision problems

1. ***Generative Models*** *is a model for generating all values for a phenomenon, both those that can be observed in the world and "target" variables that can only be computed from those observed. Simply, generative models generate both inputs and outputs, typically given some hidden parameters. More precisely, it is an approach that explicitly or implicitly model the distribution of inputs as well as outputs are known as generative models. We either*

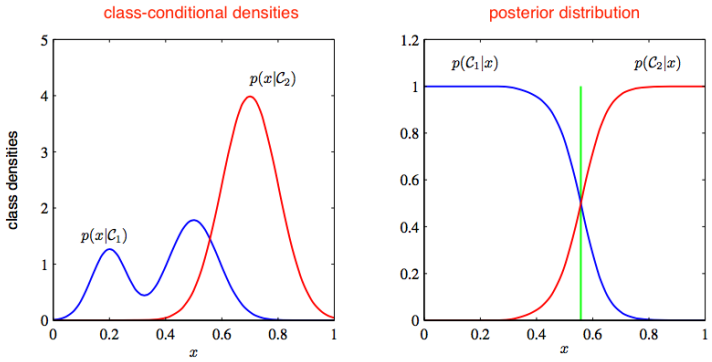
   (a) *determine* ***class-conditional densities*** $p(\mathbf{x}|\mathcal{C}_k)$*, for each class* $\mathcal{C}_k$ *individually, infer prior* $p(\mathcal{C}_k)$*, then compute* ***posterior probability*** *with Bayes rule.*

   (b) *model joint distribution* $p(\mathbf{x}, \mathcal{C}_k)$ *directly and then normalize to obtain the* ***posterior probabilities****.*

   then apply decision theory to determine class membership

2. ***Discriminative Models*** *is a model only for the target variable(s), generating them by analyzing the observed variables. In other word, approach that model the posterior probabilities* $p(\mathcal{C}_k|\mathbf{x})$ *directly. Simply, discriminative models infer outputs based on inputs.*

3. ***Discriminative Function*** *maps input* $\mathbf{x}$ *onto class label. Probabilities play no role*



Definition. ***Combining Models*** *We can combiner outputs of different model systematically using rule of probability. If distribution of input features are independent, then by conditional independence*

$$p(\mathbf{x_a}, \mathbf{x_b}|\mathcal{C}_k) = p(\mathbf{x_a}|\mathcal{C}_k)p(\mathbf{x_b}|\mathcal{C}_k)$$

Posterior of combined model

$$p(\mathcal{C}_k|\mathbf{x_a}, \mathbf{x_b}) \propto p(\mathbf{x_a}, \mathbf{x_b}|\mathcal{C}_k)p(\mathcal{C}_k) \propto \frac{p(\mathcal{C}_k|\mathbf{x_a})p(\mathcal{C}_k|\mathbf{x_b})}{p(\mathcal{C}_k)}$$

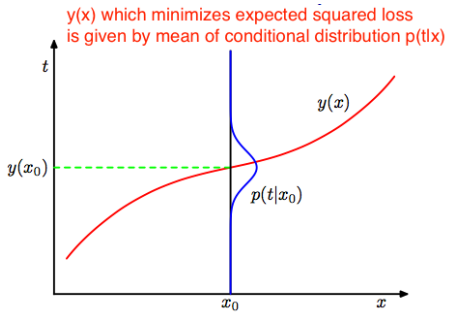
where $p(\mathcal{C}_k)$ the class prior can be estimated from the fractions of data points in each class, then normalize the resulting posterior probabilities so that they sum to one. This conditional independence assumption is an example of ***Naive Bayes Model***

Definition. ***Loss Function for Regression*** *Given predictive distribution $y(t|\mathbf{x})$, the decision theory for regression involves choosing a specific estimate $y(\mathbf{x})$, the regression function not a point prediction, of the value of target t for each input $\mathbf{x}$ such that, given loss $L(t, y(\mathbf{x}))$, the expected loss is minimized*

$$\mathbb{E}\{L\} = \iint L(t, y(\mathbf{x}))p(\mathbf{x}, t)d\mathbf{x}dt$$

If $L(t, y(\mathbf{x})) = (y(\mathbf{x}) - t)^2$, then

$$\mathbb{E}\{L\} = \iint (y(\mathbf{x}) - t)^2 p(\mathbf{x}, t)d\mathbf{x}dt \quad \rightarrow \quad y(\mathbf{x}) = \mathbb{E}_t\{t|\mathbf{x}\} \quad \textit{(by taking derivative)}$$



The regression funtion $y(\mathbf{x})$ which maximizes the expected squared loss L is given by mean of conditional distribution of $y(t|\mathbf{x})$, for a completely flexible y. ***Minkowski Loss*** *is a generalization of squared loss, whose expectation is given by*

$$\mathbb{E}\{L_q\} = \iint |y(\mathbf{x}) - t|^q p(\mathbf{x}, t)d\mathbf{x}dt$$

The minimum for $\mathbb{E}\{L_q\}$ is given by conditional mean for $q = 2$, conditional median for $q = 1$, and conditional mode for $q \rightarrow 0$

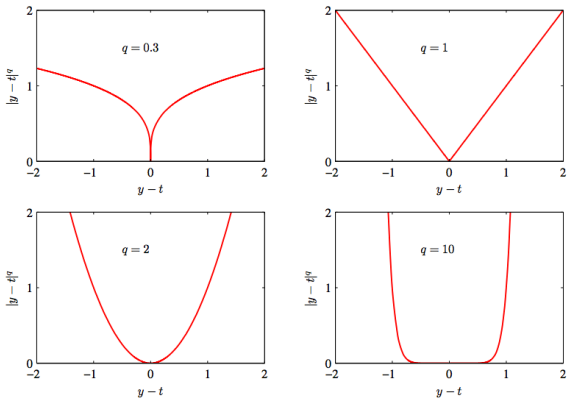


Figure 1.29 Plots of the quantity $L_q = |y - t|^q$ for various values of q.