Machine Learning

Mukesh A. Zaveri
Computer Engineering Department
Sardar Vallabhbhai National Institute of Technology, Surat
mazaveri@coed.svnit.ac.in

Machine Learning - Introduction

- data - non-numeric and unstructured (not in neat rows and columns)
- web page consists of graphs, images
- short messages, sales of product, grade points, tax assessment
- needs to handle variety of data
- words, list, images, sounds and other kinds of information
- much more than simply analyzing data
- not only histograms, averages
- range of roles, requires a range of skills
- representation, visualization, source, interpretation and understanding

Machine Learning - Introduction

- related actions, decision making, collecting, manipulating, transmitting, storing data
- organize, aggregate, visualize, present the data, decisions and negotiations
- for example, item purchase in the super store
- put in a cart, scan bar code, pay, complimentary item if any,
- stock update, manager requests for another order, special discount if any
- at end of sale or month pie charts for sale, more offer - discount, varieties

Machine Learning - Introduction

- four tasks: data architecture, data acquisition, data analysis and data archiving
- design of point of sale system, stock manager and store manager uses the same data for different purposes
- data scientist helps system architect -
  - providing input on how the data to be routed,
  - organized to support the analysis,
  - visualization and presentation of the data to the appropriate people
- data acquisition - how data collected, how the data are represented prior to analysis and presentation
- bar code representation - description of the product, price, weight, batch, packaging etc.

Machine Learning - Introduction

- data scientist actively involved in
- representing, transforming, grouping and linking the data
- are all tasks that need to occur before the data can be profitably analyzed
- analysis phase - summarization of data, inferences, visualization - animation, graph, table
- mathematical and statistical aspect
- fulfilling the needs to the data user by data scientist
- communicate results to the data user effectively whatever statistical analysis method used

Machine Learning - Introduction

- archiving of data: preservation of data in a form that makes highly reusable -
- data curation, twitter data - store tweet with a location - it may pinpoint earthquakes and tsunamis
- skills needed to do data science
- learning application domain - how data will be used in a particular context
- communicating with data users - needs and preferences of users,
- translate back and forth between technical terms of computing and statistics, vocabulary of the application domain
- able to see complex system - understanding of application domain, imagine how data will move around among all of relevant systems and people

Machine Learning - Introduction

- how data to be represented - how data can be stored and linked, metadata - data that describe how other data are arranged
- data transformation abd analysis - data available for decision making, how to transform, summarize and make interences from the data, communicate the results of analysis to users
- visualization and persentation - good way for data display, bar chart, effective means of communicating results
- attention to quality - limitations of the data, how to quantify its accuracy, able to make suggestions for improving the quality of the data in future
- ethical reasoning - important to collect, affect people's lives, ethical issues privacy, prevent misuse of data or analytical results

Machine Learning - Introduction

- great system thinker, good eye for visual displays,
- capable of thinking critically to make decisions, teamwork, different team members speccialize in different areas
- data analysis program - R, graphical user interface companion - RStudio
- real data - challenges - data not perfect
- big data is data science focused on very large data sets
- a big data problem - for example, adjusts pricing in near real time for 73 million items, based on demand and inventory
- amount of data and large amount of computations

Machine Learning - Introduction

- Clifford Stoll - cyber sleuth - Data is not information, information is not knowledge,
- knowledge is not understanding, understanding is not wisdom
- pyramid - data $\rightarrow$ information $\rightarrow$ knowledge $\rightarrow$ understanding $\rightarrow$ wisdom

Machine Learning - Introduction

- data can be used to create a model of temperature changes in different areas of the field
- this model support, improve or debunk the story
- data might be wrong, incorrect temperature data
- develop a critical approach to assess the possible situations when information might be correct or incorrect
- problem identification - look for exception cases
- statistical inference - characterize - most typical cases that occur, examine extreme cases
- for example, thunderstrom, tearing fruit off the trees, wind conditions, some trees lost more fruits or some trees less
- systematic count of lost fruit underneath a random of trees help to answer this

Machine Learning - Introduction

- exploring risk and uncertainty
- identifying the data problem to reduce uncertainty
- marketing decision, chain of events that leads to good or not good
- maximize good outcome and minimizes chance of bad one - need better decisions, needs to reduce uncertainty
- risk comes from weather, profitable or unprofitable year
- credit analysis for banking
- predict inventory, pricing inventory

Machine Learning - Introduction

- have to know data, know what you can do, know how it has to be transformed
- know how to check for problem
- think about problems in terms of data objects, procedure to process data
- follow the data - starting point of the project
- medical insurance - reimbursement, billing, procedure followed by doctor,
- chain of data consultantion, examination, test etc

Machine Learning - Introduction

- improving the efficiency of the system, auditing
- complaint with insurance records
- predict outbreak, epidemics, providing feedback to consurmer how much they pay out of pocket for various procedures
- finding out detail content, format, sender, receiver, transmission method, repositories
- user of data at each step, where data processed
- exploring data models
- data modeling - theories, strategies, tools that help in following the data by data scientist

Machine Learning - Introduction

- Ed Yourdon inroduced data flow diagram
- relational databases, enity-relationship diagram / model
- ERD describes the structure and movement of data in a system
- entity-relationship modeling at different levels
- abstract conceptual level, physical storage level etc.
- conceptual level entity is object, objects are realted by relationship
- patient and doctor are object linked by a relationship
- each object may be represented by a range of data - attributes
- patient - name, address, age
- doctor - years of experience, specialization, certifications, licenses

Machine Learning - Introduction

- for example, health care system, so many choices of designing the data
- experience and art to create workable system
- understanding current information needs and anticipating how those needs could change in future
- redesigning the system - migration, greater efficeincy, new services
- understanding and following the data with subject matter experts combined with data modeling enables data scientist to get data

Machine Learning - Introduction

- describe the sample of data we have,
- real trick is to infer what the data could mean when generalized to the larger population of data that we don't have
- key distinction between descriptive and inferential statistics
- mean - arithmetic mean - measure of central tendency
- median - another measure of central tendency
- range - measure of dispersion
- mode - another measure of central tendency
- variance - measure of dispersion
- standard deviation - another measure of dispersion - cousin to variance

Machine Learning

- searching for patterns in data
- discovery of regularities
- atomic spectra, quantum physics
- take action, classifying
- handwritten digits classification
- handcrafted rules, heuristics for distinguishing the shape of strokes
- leads to proliferation of rules and of exceptions to the rules, gives poor results
- machine learning, training set, test set
- $\mathbf{x}_1, \ldots, \mathbf{x}_N$ N training images, target vector $\mathbf{t}$
- learning funtion $\mathbf{y}(\mathbf{x})$
- once the model is trained it can then determine the identity of new digit images - test set

Machine Learning

- data generated - have regularity that we wish to learn
- individual observation is corrupted by random noise
- goal is to exploit this training set to make predictions of the value $\hat{t}$ of target variable for some new value $\hat{x}$ of input variable
- fit data using polynomial function of the form

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

- M is order of polynomial, $y(x, \mathbf{w})$ is nonlinear function of x, it is linear function of coefficients $\mathbf{w}$
- function linear in unknown parameters have important properties called linear model
- the values of coefficients determined by fitting the polynomial to the training data

Machine Learning

- by minimizing an error function that measures the misfit between the function $y(x, \mathbf{w})$, for any given value of $\mathbf{w}$ and training set data points
- error function - sum of squares of errors

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

- solve the curve fitting problem by choosing the value of $\mathbf{w}$ for which $E(\mathbf{w})$ is as small as possible
- error function is quadratic function of coefficients of $\mathbf{w}$
- its derivatives with respect to the coefficients will be linear in the elements of $\mathbf{w}$
- so minimization of error function has a unique solution denoted by $\mathbf{w}^*$ can be found in closed form
- resulting polynomial is $y(x, \mathbf{w}^*)$

Machine Learning

- problem of choosing the order M of the polynomial
- model comparison and model selection
- polynomial passes exactly through each data point and $E(\mathbf{w}^*) = 0$
- if fitted curve oscillates wildly and gives a poor representation of the function called over fitting
- goal is to achieve good generalization by making accurate predictions for new data
- root mean square RMS error $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^*)/N}$
- division by N allows us to compare different sizes of data set on an equal footing
- square root ensures that error is measured on the same scale as target variable t

Machine Learning

- small values M give large values of test set error, corresponding polynomials are inflexible and incapable of capturing oscillations in the function
- more flexible polynomials with larger values of M are becoming increasingly tuned to the random noise on the target values
- over fitting less severe as the size of the data set increases
- larger the data set, the more complex (more flexible) the model that we can afford to fit to the data
- rough heuristic number of data points should be no less than some multiple (5 or 10) of number of adaptive parameters in the model
- least square approach to find the model parameters - specific case of maximum likelihood
- overfitting can be understood as a general property of maximum likelihood

Machine Learning

- by adopting a Bayesian approach the overfitting problem can be avoided - number of parameters greatly exceeds the number of data points
- in Bayesian model the effective number of parameters adapts automatically to the size of the data set
- one technique used to control the overfitting phenomenon is that of regularization
- which involves adding a penalty term to the error function in order to discourage the coefficients from reaching large values
- simplest such penalty term takes the form of a sum of squares of all the coefficients leading to a modified error function to the form

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

$$||\mathbf{w}||^2 \equiv \mathbf{w}^T\mathbf{w} = w_0^2 + w_1^2 + \cdots + w_M^2$$

Machine Learning

- coefficient λ governs the relative importance of regularization term compared with the sum-of-squares error term
- often w_0 is omitted from regularizer because its inclusion causes the results to depend on the choice of origin for the target variable
- error function can be minimized exactly in closed form
- this techniques known as shrinkage methods because they reduce the value of the coefficients
- the particular case of quadratic regularizer is called ridge regression
- in the context of neural networks, the approach is known as weight decay
- impact of regularization term on generalization error
- λ controls the effective complexity of the model and determines the degree of overfitting

Machine Learning

- available data partitioning it into training set to determine coefficients **w** and
- validation set called hold-out set used to optimize the model complexity
- pattern recognition - there is uncertainty
- noise on measurements, through finite size of data sets
- probability theory - framework for quantification and manipulation of uncertainty
- helps in decision theory

Statistical Learning

- set of tools for understanding data
- supervised or unsupervised
- supervised learning - building a statistical model for predicting or estimating an output based on inputs
- business, medicine, astrophysics and public policy
- unsupervised learning - learn relationships and structure from input data
- for example, wages for a group of males -
- understanding association between age and education, year on his wage
- wage vs. age - wage increases with age but then decreases after age 60
- given an age predict wage from curve
- there may variability, prediction with accuracy

Statistical Learning

- wage as function of year and education
- higher education higher wage
- lower education lower wage
- better predication of wage by combining age, education and year
- regression, non-linear relationship
- predicating continuous or quantitative output value
- stock market
- categorical or qualitative output - non numerical value

Statistical Learning

- Gene expression data
- only input variables - no corresponding output
- demographic information
- which type of customers similar to other by grouping based on observed characteristics: clustering
- thousands of gene expression measurements per cell lines - hard to visualize the data
- representing 64 cell lines using two numbers z1 and z2, two principal components of data deciding the number of clusters
- relationship between gene expression levels and cancer

Statistical Learning

- Legendre and Gauss - method of least squares known as linear regression
- Fisher - linear discriminant analysis 1936
- logistic regression - 1940
- Nelder and Wedderburn - generalized linear models - 1970
- linear model - not able - fit non-linear relationship
- 1980 - nonlinear methods - Breiman, Friedman, Olshen and Stone - classification and regression trees
- modeling and prediction from data

Statistical Learning

- how to improve sales of product
- data set - sales of product in 200 market, advertising budget for each product
- control advertising expenditure, items TV, radio, newspaper etc.
- input called - predictors, independent variables, features, variables X
- output variable - called dependent variable, response Y
- observe quantitative response Y and p different variables $X = (X_1, X_2, \ldots, X_p)$
- relationship between Y and X: $Y = f(X) + \epsilon$

Statistical Learning

- f fixed but unknown function of X and ϵ random error independent of X has zero mean value
- f represents systematic information that X provides about Y
- estimate f based on observed points
- why estimate f?
- prediction and inference
- if X available, Y can not be obtained, in this case error term averages to zero
- predict $\hat{Y} = \hat{f}(X)$

Statistical Learning

- say, X characteristics of patient's blood sample - Y encoding - patient's risk for reaction to drug
- the accuracy of $\hat{Y}$ as prediction depends on two quantities: reducible error and irreducible error
- $\hat{f}$ will not be perfect estimate for f - reducible error
- possible to form a perfect estimate for f so that $\hat{Y} = f(X)$
- our prediction would still have some error in it - as Y is also function of ϵ
- variability associated with ϵ affects accuracy of prediction - irreducible error
- no matter how well estimation of f, can not reduce the error introduced by ϵ

Statistical Learning

- risk of adverse reaction might for a given patient on a given day,
- depending on manufacturing variation in the drug or
- patient's general feeling of well being on that day

$$
\begin{aligned}
E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\
&= [f(X) - \hat{f}(X)]^2 + Var(\epsilon) \\
&\quad \text{Reducible} \quad \text{Irreducible}
\end{aligned}
$$

- $E(Y - \hat{Y})^2$ - average, expected value squared difference between predicated and actual value of Y
- estimate f minimizing the reducible error
- irreducible error provides upper bound on accuracy of prediction for Y

Statistical Learning: Inference

- Y is affected as X changes
- estimate f, prediction for Y but want to understand the relationship between X and Y
- how Y changes as a function of X
- which predictors are associated with the response?
- what is relation between the response and each predictor?
- depending on the complexity of f the relationship between the response and a given predictor may also depend on the values of other predictors

Statistical Learning

- for example, company interested direct-marketing campaign
- identify individuals who respond to mailing, based on observations of demographic variables measured on each individual
- demographic variables serve as predictors and response to marketing campaign serves as outcome
- company may be interested in accurate model to predict the response using predictors
- which media contribute to sales?
- which media generate generate the biggest boost in sales?

Statistical Learning

- inference paradigm
- modeling the brand of a product that customer might purchase based on variables such as price, store location, discount levels, competition price and so forth
- how each of individual variables affects the probability of purchase?
- what effect will changing the price of a product have on sales?
- it is example of modeling for inference

Statistical Learning

- real estate setting
- relate values of homes to inputs such crime rate, zoing, distance from a river, air quality, schools, income level of community, size of houses and so forth
- how the individual input variables affect price - how much extra will a house be worth if it has a view of the river? - inference problem
- interested in predicting the value of a home given its characteristics - is under or over valued? - prediction problem
- depending on the goal is prediction, inference or a combination of the two, different methods for estimating f may be appropriate
- linear model - simple interpretable inference

Statistical Learning

- parametric or non-parametric method
- parametric method - two steps model based approach

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- assumption about functional form or shape of f, simple - linear model
- p dimensional function $f(X)$, needs to estimate $p+1$ coefficients $\beta_0, \beta_1, \ldots, \beta_p$
- fit or train the model
- least square method for fitting the model

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- disadvantage model does not match the true unknown form of f
- if chosen model is too far from the true f then estimate will be poor

Statistical Learning

- flexible models that can fit many different possible functional forms for f
- needs estimating more number of parameters
- leads to overfitting the data, follow the errors or noise too closely

$$\text{income} \approx \beta_0 + \beta_1 \times \text{education} + \beta_2 \times \text{seniority}$$

- non-parametric methods: do not make assumption about functional form of f
- estimate of f that gets as close to the data points as possible without being too rough
- have potential to accurately fit a wider range of possible shapes for f and a small number of parameters

Statistical Learning

- for example thin-plate spline used to estimate f
- provides smoothness, but problem of overfitting
- trade off between prediction accuracy and model interpretability
- linear regression simple, inflexible
- spline based flexible, overfitting
- why use more restrictive method instead of very flexible approach?
- restrictive models are much more interpretable when inference is the goal
- spline like flexible - difficult to understand how individual predictor associated with response

Statistical Learning

- flexibility and interpretability
- interested in prediction - the interpretability of predictive model is not of interest
- accurate prediction using less flexible method is possible
- supervised learning and unsupervised learning
- supervised - relate response y_i to the predictor x_i
- accurately predicting the response for future observations (prediction) or

Statistical Learning

- unsupervised learning - cluster analysis - clustering - response variable not available
- market segmentation study - customer - zip code, family income, shopping habits
- big spender and low spender
- if spending patterns available - supervised analysis
- if not, cluster the customers basis on variables measured
- groups differ with respect to some property of interest, such as spending habits
- semi supervised learning n observations, for m observations predictor and response available for $n - m$ response not available

Statistical Learning

- regression vs. classification
- variables can be characterized as quantitative or qualitative (categorical)
- quantitative - age, height, income, price of an object
- qualitative - gender, brand of object, person has debt yes or no, has cancer yes or no
- quantitative response - regression problem
- qualitative response - classification problem
- linear regression for quantitative
- logistic regression for qualitative

Statistical Learning

- no one method dominates all others over all possible data sets
- accuracy of model
- measuring the quality of fit -Ihow well predictions match the observed data

Statistical Learning

- adjusting the level of flexibility of the smoothing spline fit, produce many different fits to the data
- the degree of freedom - function of flexibility
- the degree of freedom is a quantity that summarizes the flexibility of a curve
- a more restricted curve has fewer degrees of freedom
- as model flexibility increases, training MSE will decrease but the test MSE may not
- when a given method yields a small training MSE but a large test MSE - said to be overfitting the data

Machine Learning

- random variable B - r red box and b blue box,
- random variable F - a apple and o orange
- $p(B = r) = 4/10$ $p(B = b) = 6/10$
- sum rule and product rule
- what is probability that the apple chosen?
- given chosen orange what is probability that box chosen was blue one?
- two random variables X and Y
- X take values x_i where $i = 1, \ldots, M$ and Y takes value y_j $j = 1, \ldots, L$
- consider total of N trials
- number of trials in which $X = x_i$ and $Y = y_j$ be n_{ij}

Machine Learning

- X takes value x_i number of such trials c_i
- Y takes value y_j number of trials r_j
- $$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$
- $p(X = x_i) = \frac{c_i}{N}$
- the sum of number of instance in each cell of that column $c_i = \sum_j n_{ij}$
- $p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$
- sum rule of probability, called marginal probability, summing out the other variable Y

Machine Learning

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

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \tag{1}$$
$$= p(Y = y_j | X = x_i) p(X = x_i)$$

- which is product rule of probability
- sum rule $p(X) = \sum_Y p(X, Y)$ and product rule $p(X, Y) = p(Y|X)p(X)$
- $p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$ Bayes' theorem
- plays a central role in pattern recognition and machine learning

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

Machine Learning - Regression

- predict value of one or more continuous target variables t given the value of a D dimensional vector $\mathbf{x}$ of input variables
- class of functions called linear regression models
- linear functions of adjustable parameters
- simple form linear function of input variables
- linear combinations of a fixed set of nonlinear functions of input variables known as basis functions
- such models are linear functions of parameters - simple analytical properties and yet can be nonlinear with respect to the input variables

Machine Learning - Regression

- given a training data set comprising N observations $\{\mathbf{x}_n\}$ $n = 1, \ldots, N$ together with corresponding target values $\{t_n\}$
- goal is to predict the value of t for a new value of $\mathbf{x}$
- construct function $y(\mathbf{x})$ whose values for new inputs $\mathbf{x}$ constitute the predictions for the corresponding values of t
- probabilistic perspective model predictive distribution $p(t|\mathbf{x})$
- expresses uncertainty about t for each value of $\mathbf{x}$ -
- conditional distribution makes prediction of t
- minimize expected value of chosen loss function
- significant limitations where input space of high dimensionality
- nice analytical properties

Machine Learning - Regression

- simple linear model for regression - linear combination of input variables

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D$$

- known as linear regression $\mathbf{x} = (x_1, \ldots, x_D)^T$
- it is also linear functions of parameters $w_0, \ldots, w_D$
- also linear functions of input variables
- imposes significant limitations on the model
- extend the class of models by considering
- linear combination of fixed nonlinear functions of input variables

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

- $\phi_j(\mathbf{x})$ known as basis functions

Machine Learning - Regression

- total number of parameters in the model will be M
- w_0 parameter allows fixed offset in the data called bias parameters
- dummy bias function $\phi_0(\mathbf{x}) = 1$

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

- called linear models, function is linear in $\mathbf{w}$
- linearity in parameters greatly simplify the analysis of this class of models
- $\mathbf{w} = (w_0, \ldots, w_{M-1})^T$ and $\phi = (\phi_0, \ldots, \phi_{M-1})^T$
- many applications apply some form of fixed pre-processing or feature extraction to original data variables
- original variables comprise vector $\mathbf{x}$ then features expressed in terms of basis functions $\{\phi_j(\mathbf{x})\}$

Machine Learning - Regression

- using nonlinear basis functions, $y(\mathbf{x}, \mathbf{w})$ to be nonlinear function of input vector $\mathbf{x}$)
- polynomial regression single input variable x, basis function take the form of powers of x $\phi_j(x) = x^j$
- limitation of polynomial basis functions is that they are global functions of the input variables so the changes in one region of input space affect all other regions
- this can be resolved by dividing the input space up into regions and fitting a different polynomial in each region leading to spline functions
- other choices for basis

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$

Machine Learning - Regression

- μ_j govern locations of basis functions in input space and parameter s governs their spatial scale
- referred Gaussian basis functions
- basis function multiplied by adaptive parameters w_j
- sigmoidal basis function

$$\phi_j(x) = \sigma\left(-\frac{x-\mu_j}{s}\right)$$

- $\sigma(a)$ is logistic sigmoid function

$$\sigma(a) = \frac{1}{1+\exp(-a)}$$

Machine Learning - Regression

- tanh function can be used - related to logistic sigmoid
  $\tanh(a) = 2\sigma(2a) - 1$
- general linear combination of logistic sigmoid functions is equivalent to general linear combination of tanh functions
- another choice is Fourier basis - expansion in sinusoidal functions
- each basis function represents a specific frequency and has infinite spatial extent
- by contrast, basis functions that are localized to finite regions of input space necessarily comprise a spectrum of different spatial frequencies
- basis functions localized in both space and frequency - wavelets - mutually orthogonal
- simple case $\varphi(\mathbf{x})$ of basis functions - identity $\phi(\mathbf{x}) = \mathbf{x}$

Machine Learning - Regression

- least square and maximum likelihood
- target variable t given by deterministic function $y(\mathbf{x}, \mathbf{w})$ with additive Gaussian noise

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon$$

- ϵ zero mean Gaussian random variable with precision β (inverse variance)

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

- assume if squared loss function then optimal prediction for a new value $\mathbf{x}$ will be given by conditional mean of target variable
- conditional mean

$$\mathbb{E}[t|\mathbf{x}] = \int tp(t|\mathbf{x})dt = y(\mathbf{x}, \mathbf{w})$$

- Gaussian noise assumption implies that the conditional distribution of t given $\mathbf{x}$ is unimodal, which may be inappropriate for some applications

Machine Learning - Regression

- multivariate target $\mathbf{t}$ - grouping target variables $\{t_n\}$
- a data set of inputs $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and corresponding values $t_1, \ldots, t_N$

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\phi(\mathbf{x}_n), \beta^{-1})$$

- $\mathbf{x}$ always appears in set of conditioning variables

$$\begin{aligned} \ln p(\mathbf{t}|\mathbf{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^T\phi(\mathbf{x}_n), \beta^{-1}) \\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w}) \end{aligned} \tag{2}$$

Machine Learning - Regression

- sum of squares error function

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2$$

- maximum likelihood to determine $\mathbf{w}$ and β
- maximization of likelihood function under a conditional Gaussian noise distribution for a linear model is equivalent to minimizing a sum of squares error function given by $E_D(\mathbf{w})$
- gradient of log likelihood function

$$\nabla \ln p(\mathbf{t}|\mathbf{w}, \beta) = \beta\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}\phi(\mathbf{x}_n)^T$$

setting gradient to zero

Machine Learning - Regression

$$0 = \sum_{n=1}^{N} t_n \phi(\mathbf{x}_n)^T - \mathbf{w}^T \left(\sum_{n=1}^{N} \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^T \right)$$

solving for **w**

$$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{\Phi}^T \mathbf{\Phi} \right)^{-1} \mathbf{\Phi}^T \mathbf{t}$$

- known as normal equations for least squares problem
- $\mathbf{\Phi}$ is $N \times M$ matrix called design matrix elements given by $\Phi_{nj} = \phi)j(\mathbf{x}_n)$

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

Machine Learning - Regression

$$\mathbf{\Phi}^{\dagger} \equiv (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T$$

- known as Moore Penrose pseudo inverse of matrix $\mathbf{\Phi}$
- bias parameter w_0

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - w_0 - \sum_{j=1} M - 1 w_j \phi_j(\mathbf{x}_n)\}^2$$

derivative with respect to w_0 equal to zero, solving for w_0

$$w_0 = \bar{t} - \sum_{j=1} M - 1 w_j \bar{\phi}_j$$

$$\bar{t} = \frac{1}{N} \sum_{n=1}^{N} t_n \quad \bar{\phi}_j = \frac{1}{N} \sum_{n=1}^{N} \phi_j(\mathbf{x}_n)$$