

CS 480 Course Notes

Introduction to Machine Learning

Michael Socha

University of Waterloo
Winter 2019

Contents

1	Course Overview	1
2	Introduction - What is Machine Learning?	2
2.1	Learning Frameworks	2
2.2	Challenges	2
3	Decision Trees	3
3.1	Predictions Using Decision Trees	3
3.2	Encoding Functions in Decision Trees	3
3.3	Training and Testing	3
3.3.1	Information Content	4
3.3.2	Entropy	4
3.4	Overfitting	4
3.5	Inductive Bias	4
3.6	Advantages and Disadvantages	4
4	Evaluation of Learning	6
4.1	Performance Formulation	6
4.2	Common Learning Challenges	6
4.3	Training vs Validation vs Test Sets	6
4.4	Bias-Variance Decomposition	6
4.5	Cross Validation	7
4.6	Bootstrapping	7
4.7	Avoiding Overfitting	7
4.8	Performance Evaluation of Classifiers	8
4.8.1	Accuracy and Error	8
4.8.2	Precision and Recall	8
4.8.3	F-Measure	8
4.8.4	Sensitivity and Specificity	9
5	Instance-Based Learning	10
5.1	Parametric vs Non-Parametric Methods for Supervised Learning	10
5.1.1	Approximation	10
5.1.2	Efficiency	10
5.2	K-Nearest Neighbors	10
5.2.1	Implementation	10
5.2.2	Distance Function	10
5.2.3	Decision Boundaries	11
5.2.4	Selection of K	11
5.2.5	Pre-Processing	11
5.2.6	Distance-Weighted Nearest Neighbor	11
5.2.7	High Dimensionality	12
6	Perceptron	13

6.1	Formulation	13
6.2	Decision Boundary	13
6.3	Learning Algorithm	13
6.3.1	Hyper-parameters	14
6.3.2	Extensions	14
6.4	Convergence	14
6.4.1	Rosenblatt's Perceptron Convergence Theorem	14
6.5	Limitations	15
7	Linear Models	16
7.1	Formulation	16
7.2	Training Parameters	16
7.2.1	Gradient Descent Solution	16
7.2.2	Closed Form Solution	17
7.3	Regularization	17
7.3.1	P-norm	17
7.3.2	Ridge Regression	18
7.3.3	Lasso Regression	18
7.4	Application to Classification Problems	18
8	Probabilistic Methods	19
8.1	Bayesian Learning	19
8.2	Maximum a Posteriori	20
8.3	Maximum Likelihood	20
8.4	Determining Priors	20
8.5	Bayesian Linear Regression	20
9	Linear Models for Classification	22
9.1	Probabilistic Framework for Classification	22
9.2	Generative Learning	22
9.2.1	Training	22
9.2.2	Decision Boundary	22
9.2.3	Laplace Smoothing	22
9.2.4	Multiclass Classification	23
9.3	Discriminative Learning	23
9.3.1	Training	23
9.3.2	Decision Boundary	23
10	Support Vector Machines	24
10.1	Margins	24
10.1.1	Functional Margins	24
10.1.2	Geometric Margins	24
10.1.3	Margin of Training Set	24
10.2	Optimal Margin Classifier	24
10.3	Constrained Optimization	24

10.3.1	Lagrange Multipliers	24
10.3.2	Primal and Dual Problem	25
10.4	Training	25
10.5	Handling Non-Linearly Separable Data	26
11	Kernel Methods	27
11.1	Motivation	27
11.2	Common Kernels	27
11.3	Kernel Properties	27
11.4	Kernel Perceptron Algorithm	28

1 Course Overview

This is an applied introductory machine learning course covering the basics of machine learning algorithms and data analysis. Topics covered include:

- Regression analysis
- Probabilistic modeling
- Support vector machines
- Supervised vs unsupervised learning
- Reinforcement learning
- Neural networks

2 Introduction - What is Machine Learning?

Machine learning is the field of study of how computers can improve their performance at tasks (i.e. learn) without being explicitly programmed to do so. Machine learning can be useful for tasks for which it is difficult to write a step-by-step imperative program. Sample applications include optical character recognition, computer vision, and game playing.

2.1 Learning Frameworks

- **Supervised Learning:** Goal is to learn a function based on its input and output (e.g. determining if an email is spam based on a set of emails labeled as spam or not spam).
- **Unsupervised Learning:** Goal is to learn a function based on its input alone (e.g. organizing data into clusters).
- **Reinforcement Learning:** Goal is to learn a sequence of actions that maximize some notion of reward (e.g. learning how to control a vehicle to perform some maneuver).

2.2 Challenges

Some of the challenges facing machine learning today are:

- Dealing with large amounts of data (algorithm complexity and distributed computing become very relevant)
- Generating reproducible results
- Challenges concerning real-world adoption of work (e.g. human computer interaction, robustness, ethical concerns)

3 Decision Trees

Decision trees contain questions as nodes and answers as edges, and serve to guide to an answer based on a set of observations.

3.1 Predictions Using Decision Trees

Consider a data set of employees and their job satisfaction, where we use the data set to predict whether an employee is satisfied with their job. An employee can have many features, such as age, salary, seniority, working hours. It is infeasible to build all possible decision trees when there are many features due to the exponential growth of the tree. Instead, for predictive purposes, it makes sense to focus on those questions that are informative to the prediction. For example, if there is no major difference in job satisfaction based on an employee's age, then it is unnecessary to include age in a decision tree, since the answer is not informative for the prediction.

In general, supervised learning problems have a set of possible inputs X , an unknown target function $f : X \rightarrow Y$, and a set of function hypotheses $\{h|h : X \rightarrow Y\}$. Supervised learning algorithms accept training examples (a pair (x, y) where $x \in X, y \in Y$) and output a function hypotheses that approximates f . When using decision trees for learning, each decision tree is a function hypothesis.

3.2 Encoding Functions in Decision Trees

Boolean functions can be fully expressed in decision trees, with one branch for true and another for false. Other function can be approximated as a boolean function. For example, instead of a node asking what an employee's salary is, it could ask whether the salary is above a certain amount.

3.3 Training and Testing

The key idea behind decision tree generation algorithms is to grow a tree until it correctly classifies all training examples. The rough procedure followed is:

1. If all training data has the same class, create a leaf node and return.
2. Else, create the best (i.e. most informative) node on which to split the data.
3. Split the training set over the above node.
4. Continue the procedure on each subset of training data generated.

3.3.1 Information Content

Criteria for finding the “best” split of data can be defined mathematically. If event E occurs with probability $P(E)$, then when E occurs, we receive $I(E) = \log_2 \frac{1}{P(E)}$ bits of information. This can be interpreted as that less likely events yield more information.

3.3.2 Entropy

An information source S which emits results s_1, s_2, \dots, s_i with probabilities p_1, p_2, \dots, p_i produces information at $H(S) = \sum_i p_i I(s_i)$. $H(S)$ is known as the information entropy of S . Information entropy can vary between 0, which indicates no uncertainty (i.e. all members of S in same class), and 1, which indicates high uncertainty (i.e. equal probability of all classes). The best split of data for classification is one that maximally decreases its entropy (a concept known as information gain).

3.4 Overfitting

A hypothesis $h_1 \in H$ is said to overfit training data if there is some alternative hypothesis $h_2 \in H$ that has a larger error over the training data but a smaller error over a larger set of inputs. Overfitting can occur due to errors in a data set or just due to coincidental irregularities (especially in a small dataset). Overfitting can be avoided by removing nodes with low information gain, either by stopping decision tree construction early or by pruning such nodes after the tree is constructed.

3.5 Inductive Bias

Inductive bias refers to the assumptions made about the target function to predict future outputs. Common inductive biases for decision trees are:

- Assumption that simplest hypothesis is the best (i.e. Occam’s razor).
- Decision trees with information gain closer to the root are considered better.

These are examples of preference bias, which influence the ordering of the hypothesis space. This is distinct from restriction bias, which limits the hypothesis space.

3.6 Advantages and Disadvantages

Decision trees are good for:

- Ease of interpretation
- Speed of learning

Limitations of decision trees include:

- High sensitivity, with tree output changing significantly due to small changes in input.
- Not good for learning data sets without axis-orthogonal (i.e. constant in at least 1 dimension) decision boundaries.

4 Evaluation of Learning

4.1 Performance Formulation

Let \hat{y} be an output generated by a function f approximating some target function. Let y be the corresponding output of the target function. A loss function $l(y, \hat{y})$ can be used to measure the accuracy of the approximation function f . Some common loss functions include:

- Squared Loss: $l(y, \hat{y}) = (y - \hat{y})^2$
- Absolute Loss: $l(y, \hat{y}) = |y - \hat{y}|$
- Zero/One Loss: $l(y, \hat{y}) = 1_{y \neq \hat{y}}$

We assume that the data coming from our target function comes from some probability distribution D , and that our training data is a random sample of (x, y) pairs from D . A Bayes Optimal Classifier is a classifier that for any input x , returns the y most likely to be generated by D .

Based on the available training data, the goal of supervised learning is to find a mapping f from x to y such that generalization error $\sum_{(x,y)} D(x, y) l(y, f(x))$ is minimized. However, since D is unknown, we instead estimate the error from the average error in our training or test data, which is $\frac{1}{N} \sum_{n=1}^N l(y_n, f(x_n))$.

4.2 Common Learning Challenges

Some common challenges in learning include:

- Inductive bias of the algorithm being distant from the concept actually being learned.
- The data itself having challenging characteristics, such as lots of noise, ambiguity, or missing information.

4.3 Training vs Validation vs Test Sets

Models are initially built based on a training dataset. Test sets (also known as holdout sets) are then used to estimate the generalization error. Validation sets are also used to measure the model's performance, but unlike test sets, validation sets can make changes to the model's parameters.

4.4 Bias-Variance Decomposition

Many machine learning algorithms are based on building a formal model based on the training data (e.g. a decision tree). Models have parameters, which are characteristics that can help

in classification (e.g. a node in a decision tree). Models may also have hyper-parameters, which in turn control other parameters in a model (e.g. max height of decision tree).

Generalization errors result from a combination of noise, variance, and bias. Bias concerns how well the type of model fits the data. Models with high bias pay little attention to training data and suffer from underfitting, while models with low bias may pay too much attention to training data and become overfitted. Bias and variance tend to be at odds with one another (high bias typically leads to low variance, and vice versa). For example, a decision tree that makes the same prediction for all input has high bias and low variance, while a decision tree trained to return a correct prediction for each point of training data will have low bias and likely high variance.

4.5 Cross Validation

Cross validation is a technique for measuring how well a model generalizes. The idea behind it is to break up a training data set into K equally sized partitions, and use $K - 1$ of the partitions as training data and the remaining partition for testing. This should be repeated K times, so that all points of data are at some point used for testing. Higher values of K lower the amount of variance of in the error estimation. To avoid training and testing data having a different probability distribution, the data should be shuffled before being split.

4.6 Bootstrapping

Bootstrapping is an alternative to cross validation where instead of dividing a training data set into partitions, a random sample of points (with possible duplicates) is used as training data. The remaining points are then used as testing data, with the goal being similar to that of cross validation.

4.7 Avoiding Overfitting

Overfitting often occurs when training data has many features, which allows for an approximation of the target function with many degrees of freedom. Overfitting can be avoided by only considering features with a strong correlation to the output. Cross-validation could be applied as follows:

1. Divide training data into K groups at random.
2. Within each group, find a small set of features with strong correlation to the output. Running this step for each group individually instead of for the entire training set further helps avoid overfitting.
3. Build a classifier using the features and examples from the $K - 1$ groups.
4. Use the classifier to predict the examples and in group K and measure the error.

5. Repeat 2-4 to produce an overall cross-validation estimate of the error.

Bootstrapping can be applied in a similar way to avoid overfitting.

4.8 Performance Evaluation of Classifiers

Consider the following terminology for classification problems:

- True positive (TP) - Examples of class 1 predicted as class 1
- False positive (FP) - Examples of class 0 predicted as class 1 (Type 1 Error)
- True negative (TN) - Examples of class 0 predicted as class 0
- False negative (FN) - Examples of class 1 predicted as class 0 (Type 2 Error)

4.8.1 Accuracy and Error

The following formulas can be used to measure accuracy and error:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$ErrorRate = \frac{FP + FN}{TP + TN + FP + FN}$$

4.8.2 Precision and Recall

Precision and recall can be measured as follows:

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN}$$

Precision measures the ratio of positive predictions that were correct, while recall measures the ratio of total positive instances that were predicted. Similarly to how variance and bias are often at odds with one another, so are precision and recall.

4.8.3 F-Measure

An F-measure (also known as a F1 score) measures a model's accuracy by taking into account both precision and recall as follows:

$$F = \frac{2PR}{P + R}$$

To adjust the relative importance of precision vs recall, a weighted F-measure can be used, which is defined as follows:

$$F = \frac{(1 + \beta^2)PR}{\beta^2P + R}$$

In a standard F-measure, $\beta = 1$, $\beta < 1$ means that precision is valued over recall, while $\beta > 1$ means recall is valued over precision.

4.8.4 Sensitivity and Specificity

Sensitivity is the same measure as recall. Specificity is a measure of how well a classifier avoids false positives, and is measured as:

$$Specificity = \frac{TN}{TN + FP}$$

5 Instance-Based Learning

5.1 Parametric vs Non-Parametric Methods for Supervised Learning

Datasets can be represented as a set of points in a high-dimensional space; a data point with n features f_1, f_2, \dots, f_n can be represented with the feature vector (f_1, f_2, \dots, f_n) in n -dimensional space. Parametric methods of supervised learning attempt to model the data using these parameters, while non-parametric (also known as instance-based) methods do not.

5.1.1 Approximation

Parametric methods use parameters to create global approximations. Non-parametric methods instead create approximations based on local data.

5.1.2 Efficiency

Parametric methods do most of their computation beforehand, and then summarize their results in a set of parameters. Non-parametric methods tend to have a shorter training time but a longer query answering time.

5.2 K-Nearest Neighbors

K-nearest neighbors (KNN) is a common non-parametric method. The idea is to predict the value of a new point based on the values of the K most similar (i.e. closest) points.

5.2.1 Implementation

A common implementation of KNN involves looping through all N points in a training set and computing their distance to some point x . Then the K nearest points are selected. This process can be sped up by storing the data points in a data structure that helps facilitate distance-based search (e.g. a k-d tree).

5.2.2 Distance Function

“Nearby” means of minimal distance, which is commonly defined by Euclidean distance. Other distance functions $d(x, x')$ can be used, though must meet the following conditions:

- $d(x, x') = d(x', x)$ (i.e. symmetric)

- $d(x, x) = 0$ (i.e. definite)
- $d(a, c) \leq d(a, b) + d(b, c)$ (i.e. triangle inequality holds)

5.2.3 Decision Boundaries

Decision boundaries define the borders of a single classification of input. These boundaries are formed of sections of straight lines that are equidistant to two points of different classes. A highly jagged line is an indicator of overfitting, while a simple line is an indicator of underfitting.

5.2.4 Selection of K

The selection of the value of K is a bias-variance tradeoff. Low values of K have high variance but low bias, while high values of K have low variance but high bias. High-values of K result in smoother decision boundaries, which can be a sign of underfitting, and vice versa.

K can be selected experimentally by evaluating the performance for different values of K through cross-validation or against a testing set. In theory, as the number of training examples approaches infinity, the error rate of a 1NN classifier is at worst twice that of the Bayes Optimal Classifier.

5.2.5 Pre-Processing

Some common forms of pre-processing for KNN include:

- Removing undesirable inputs. Common removal methods are:
 - Editing methods, which involve eliminating noisy points of data.
 - Condensation methods, which involve selecting a subset of data that produces the same or very similar classifications.
- Use custom weights for each feature (not all features may be equally relevant for the situation)

5.2.6 Distance-Weighted Nearest Neighbor

A common problem with KNN is that it can be sensitive to small changes in the training data. One way to mitigate with drawback is to compute a weight for each neighbor based on its distance (e.g. through a Gaussian distribution), and this weight determines how much of an influence that point's value has. This differs from standard KNN which weighs the values of the K nearest neighbors equally and ignores all other values.

5.2.7 High Dimensionality

In uniformly distributed high-dimensional spaces, distances between points tend to be roughly equal, since there are so many features that changing a few features results in only a small change in distance. However, KNN can still be applied in practice for high-dimensional spaces, since data in high-dimensional spaces tends to be concentrated around certain hubs rather than uniformly distributed.

6 Perceptron

A perceptron is a binary classification algorithm that accepts a set of features associated with weights.

6.1 Formulation

Let $x = (x_1, x_2, \dots, x_D)$ be a feature vector with D features, let $w = (w_1, w_2, \dots, w_D)$ be the corresponding weight vector, and let b be some fixed bias/threshold term. Activation a is defined as:

$$a = w^T x + b = \sum_{d=1}^D w_d x_d + b$$

If $a \geq 0$, then the perceptron classifier predicts a positive classification, and otherwise it predicts a negative classification.

6.2 Decision Boundary

A perceptron decision boundary is where the sign of the activation changes from negative to non-negative. It can be expressed as $a = w^T x + b = 0$, so $w^T x = -b$. Hence, the decision boundary is a hyperplane perpendicular to w . The bias term can shift the hyperplane, but does not change its alignment.

If the weight and bias terms are scaled together, then the decision boundary does not change. However, it is common to normalize the weight vector to have a magnitude of 1.

6.3 Learning Algorithm

The perceptron learning algorithm starts with any (e.g random) w and b parameters, and iterates over each point training point as follows:

1. Compute the activation a for the point.
2. Compute the classification y based on the activation a (i.e. 1 is $a \geq 0$, otherwise 0).
3. If the perceptron correctly classifies the point, continue.
4. Otherwise, update the weights as $w_d = w_d + yx_d$, and update the bias as $b = b + y$.
This guarantees that this point is correctly classified for this iteration.

The perceptron algorithm is guaranteed to converge on parameters (i.e. weights and bias) that correctly classify a set of points if such parameters exist.

The algorithm is considered to be:

- **Online**, since data becomes available in sequential order instead of all at once.
- **Error driven**, since it requires an incorrectly classified point to update its parameters.

6.3.1 Hyper-parameters

The only hyper-parameter of the perceptron algorithm is the maximum number of iterations of the above algorithm (i.e. the number of times each point is considered). A high number of iterations can lead to overfitting, while a low number of iterations can lead to underfitting. Since the order in which training points are encountered matters in the training algorithm, it is important to permute the order of examples between iterations.

6.3.2 Extensions

A potential problem with the perceptron algorithm described above is that later training points are weighted more than earlier ones. For instance, it is possible that after selecting weights that correctly classify thousands of training points, a single mislabeled point updates the weights in a way that drastically changes the decision boundary in order to correctly classify this new point, but in doing so, mislabels many training points that were previously correctly classified.

This problem can be mitigated by favoring weight vectors that “survive” a long time. This can be done by letting the activation a use a weighted sum of weight vectors, each of which in turn is weighted by its survival time (i.e. voted perceptron algorithm). Alternatively, a single weight vector and bias term can be tracked using the weighted average of their survival times (i.e. average perceptron algorithm).

6.4 Convergence

Data can only be classified by a perceptron if it is linearly separable. Otherwise, the perceptron algorithm will not converge, and the decision boundary will oscillate.

The speed of convergence of the perceptron algorithm depends on its margin, which is the distance from the separating hyperplane to the nearest data point. Large margins result in faster convergence, since there is more “wiggle room” to adjust the parameters defining the hyperplane.

6.4.1 Rosenblatt’s Perceptron Convergence Theorem

Let the margin γ of a linearly separable dataset D be the largest attainable margin on that dataset. Rosenblatt’s perceptron convergence theorem states that if $||x|| < 1$ for all $x \in D$, then the perceptron algorithm will converge after at most $\frac{1}{\gamma^2}$ updates. This is an upper bound; the number of updates also depends on the dataset and the initial parameters.

6.5 Limitations

Two key limitations of perceptron classifiers are:

- The generated solutions (i.e. separating hyperplanes) may not be unique.
- Non linearly-separable data cannot be modeled. In some cases, this limitation can be overcome by transforming feature vectors so they become linearly separable, or by combining multiple perceptrons, which create a network that might be able to learn non-linear boundaries.

7 Linear Models

7.1 Formulation

A loss function concerning the distance d between the value of points x predicted by some model w, b and their actual values y can be minimized as follows:

$$\min_{w,b} \sum_n d(y_n, w^T x_n + b) + \lambda R(w, b)$$

The first term measures the training error, while the second term is known as a regularizer, and exists to prevent overfitting. λ is a constant hyper-parameter while R is a function that imposes some penalty based on the complexity of parameters w, b .

7.2 Training Parameters

7.2.1 Gradient Descent Solution

The gradient of error can be used to determine the direction of steepest decline in error, and the training algorithm then adjusts parameters to move in that direction. The training algorithm maintains the state of some parameter z , finds its gradient g , and updates z to be $z + \eta g$, where η is the learning rate (i.e. the size of each step).

As an example, consider a loss function:

$$L(w, b) = \sum_n \exp(-y_n(w^T x_n + b)) + \frac{\lambda}{2} \|w\|^2$$

The partial derivative of L over b is $-\sum_n y_n \exp(-y_n(w^T x_n + b))$, so b is adjusted in each step to $b - \eta \sum_n y_n \exp(-y_n(w^T x_n + b))$.

The partial derivative of L over w is $(-\sum_n y_n x_n \exp(-y_n(w^T x_n + b))) + \lambda w$, so w is adjusted in each step to $w - \eta((-\sum_n y_n x_n \exp(-y_n(w^T x_n + b))) + \lambda w)$.

The learning rate η impacts the size of each step. A high learning rate can result in oscillations around local minima, while a low learning rate results in training being slow. It is common to lower the learning rate during later steps. When parameters start to not change by much, then a local minimum has likely been approached, and the algorithm can stop.

Note that gradient descent does not guarantee convergence to a global minima; it can get stuck in a local minima that may not be optimal. However, approximations using linear functions have a single global minimum.

7.2.2 Closed Form Solution

The gradient descent solution iteratively approaches some minimum. On the other hand, a closed form solution provides a formula for the absolute minimum. A closed form solution is not always attainable, but this section focuses on an attainable case, which is where the loss function is measured by Euclidean distance.

Our model can start as linear; we assume that the target function can be modeled as $f_w(x) = w_0 + w_1x_1 + w_2x_2 + \dots + w_Dx_D$. A linear model is often a poor approximation, so further input variables can be added based on x_1, \dots, x_D , such as:

- Transformations of variables, such as $\log x_i$
- Interaction terms, such as $x_i \cdot x_{i+1}$
- Numeric encoding of qualitative variables

The goal of the closed-form solution is to minimize $\sum_{i=1}^n (y_i - w^T x_i)^2$ (i.e. minimize Euclidean distance of errors). It is convenient to rewrite the problem in matrix format where:

- X is a $N \times D$ training data matrix, where $X_{n,d}$ is the value of feature d on example n .
- w is a size D column vector storing weights.
- $\hat{Y} = Xw$ are the predicted values.

Since we aim to minimize squared error, we aim to minimize $\|\hat{Y} - Y\|^2 + \lambda\|w\|^2$. Taking the derivative of this against w and setting it to 0 results in $w = (X^T X + \lambda I_D)^{-1} X^T Y$. This equation takes $\theta(D^3 + D^2N + DN)$ steps to compute, so this method can be computationally expensive for very high-dimensional problems.

7.3 Regularization

The Gauss-Markov theorem states that in a linear regression model, such as the one described above, the least-squares estimate (i.e. Euclidean distance) results in a model with the least bias. We nonetheless add a regularization term, which increases bias, to avoid overfitting. A few desirable features of a regularization function are:

- It is convex
- Small (even 0) weight vector components are preferred (note that this is a form of inductive bias).

7.3.1 P-norm

The p-norm of the weight vector can be used as a regularization function, defined as follows:

$$R(w) = (\sum_d |w_d|^p)^{\frac{1}{p}}$$

The p-2 norm is the same as the Euclidean norm used above.

7.3.2 Ridge Regression

Ridge regression uses the following regularization function:

$$R(w) = \sum_d w_d^2$$

Ridge regression tends to give a smooth solution with low weights.

7.3.3 Lasso Regression

Lasso regression uses the following regularization function:

$$R(w) = \sum_d |w_d|$$

Lasso regression leads to a less smooth and more computationally expensive solution than ridge regression. It is equivalent to a p-1 norm.

7.4 Application to Classification Problems

The perceptron algorithm covered above can find a hyperplane that splits a linearly separable dataset. However, the perceptron algorithm does not converge for datasets that are not linearly separable.

Due to this limitation, it can make sense to relax the problem to finding a hyperplane that splits data with the fewest possible classification errors. However, due to the discrete nature of 0/1 error, it is NP hard to minimize. Thus, 0/1 classification functions are commonly modeled with continuous approximations. One common class of approximations are convex surrogate loss functions, which extend between 0 and 1 and are convex, which is a property that makes them easy to minimize.

8 Probabilistic Methods

Data is often incomplete, indirect, or noisy. Considering such forms of uncertainty can help build better models.

8.1 Bayesian Learning

Bayes' theorem describes the probability of an event H given evidence e .

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)}$$

The terms of the above equation are known as follows:

- $P(H|e)$: Posterior probability
- $P(e|H)$: Likelihood
- $P(H)$: Prior probability
- $P(e)$: Normalizing constant

Bayesian learning consists of determining the posterior probability using Bayes' theorem.

As an example, consider a bag of candy with an unknown flavor ratio. The possible flavors are cherry and lime. The hypotheses H could be:

- h_1 : 100% cherry, 0% lime, $P(h_1) = 0.1$
- h_2 : 75% cherry, 25% lime, $P(h_2) = 0.2$
- h_3 : 50% cherry, 50% lime, $P(h_3) = 0.4$
- h_4 : 25% cherry, 75% lime, $P(h_4) = 0.2$
- h_5 : 0% cherry, 100% lime, $P(h_5) = 0.1$

Sample examples E could be:

- e_1 : 1st candy is cherry
- e_2 : 2nd candy is lime
- e_3 : 3rd candy is lime

Suppose of the first 10 candies, all are lime, and let us compute $P(h_4|e)$. $P(e|h_4) = 0.75^{10}$. $P(e) = \sum_i P(e|h_i)P(h_i) \approx 0.1114$. It follows that $P(h_4|e) = \frac{P(e|h_4)P(h_4)}{P(e)} = \frac{(0.75)^{10} \cdot 0.2}{0.1114} \approx 0.0987$. The probability of the other hypothesis can also be calculated in this manner.

Bayesian probability is:

- **Optimal**; given a prior probability, no prediction is correct more often than the Bayesian prediction.
- **Overfitting-free**; all hypothesis are weighted and considered, eliminating overfitting.

However, Bayesian learning can be intractable when the hypothesis space grows very large, often as a result of approximating a continuous hypothesis space with many discrete hypotheses.

8.2 Maximum a Posteriori

Maximum a Posteriori (MAP) makes predictions based on only the most probable hypothesis h_{MAP} (i.e. $h_{MAP} = \arg \max_h P(h|e)$). This differs from Bayesian learning, which makes predictions for all hypotheses weighted by their probability. However, MAP and Bayesian predictions tend to converge as the amount of data increases, and overfitting can be mitigated by giving complex hypotheses a low prior probability. However, finding h_{MAP} can be difficult or intractable.

8.3 Maximum Likelihood

Maximum Likelihood (ML) simplifies MAP by assuming uniform prior probabilities, and then makes a prediction based on the most probable hypothesis h_{ML} . ML tends to be less accurate than MAP and Bayesian predictions, and due to prior probabilities being uniform, is subject to overfitting. However, finding h_{ML} is often easier than h_{MAP} , since finding h_{ML} for $P(e|h)$ is equivalent to calculating it for $\sum_n \log P(e_n|h)$, which is often easier to optimize.

8.4 Determining Priors

Selecting a prior is important for Bayesian learning and MAP estimates. If the posterior distributions $P(H|x)$ are in the same family of functions as the prior probability distributions $P(H)$, then the prior and posterior are known as conjugate distributions, and have properties that make them easy to work with.

8.5 Bayesian Linear Regression

The linear regression problem discussed in the above section can be reformulated by assuming part of y to be set from a probability distribution $\eta \sim N(0, \omega^2)$ as follows:

$$y = w^T b(x) + \eta$$

b is a fixed set of basis functions (i.e. $b(x) = [b_1(x), b_2(x), \dots, b_M(x)]$).

The goal of Bayesian linear regression is not to find the best model parameter w , but rather to find a model parameter's posterior distribution as follows:

$$P(w|y, X) = \frac{P(y|w, X)P(w|X)}{P(y|x)}$$

9 Linear Models for Classification

9.1 Probabilistic Framework for Classification

As covered in the above section, learning can be viewed as a problem of statistical inference. If we have access to the underlying probability distribution of a set of data, then we can form an optimal Bayesian classifier. In practice, we typically do not know the underlying probability distributions, so we have to estimate them from the available training data.

It is usually best to choose a family of parametric distributions (e.g. Gaussian or Binomial) and then determine which parameters describe the available training data the best. This is known as a density estimate, and we assume that each point of training data is independently selected from the same distribution.

9.2 Generative Learning

The idea behind generative learning is to separately model $P(x|y)$ and $P(y)$, and to use Bayes' theorem to estimate $P(y|x)$.

$P(x|y) = P(x_1, x_2, \dots, x_m|y)$. Using what is known as the naive Bayes assumption, we assume all points x_i are conditionally independent. Thus, $P(x|y) = P(x_1|y) \cdot P(x_2|y) \cdot \dots \cdot P(x_m|y)$, greatly simplifying the probability calculation.

9.2.1 Training

Training involves finding parameters that maximize $P(y|x) \propto P(y)P(x|y) = \prod_{i=1}^n (P(y_i) \prod_{j=1}^m P(x_{i,j}|y_i))$. The derivative of the log of the above expression can be taken against output variables to find the probabilities of each output. For example, suppose θ_1 is the likelihood that the output of a binary classification problem is 1. The above expression is maximized at $\theta_1 = \frac{1}{n} \sum_{i=1}^n y_i$, which is the number of outputs where $y = 1$ over the total number of outputs.

9.2.2 Decision Boundary

The decision boundary occurs where $\frac{P(y=1|x)}{P(y=0|x)} = 1$, or equivalently, the log-likelihood ratio $\log \frac{P(y=1|x)}{P(y=0|x)} = 0$.

9.2.3 Laplace Smoothing

Laplace smoothing, also known as additive smoothing, is a technique used to smooth categorical data. For instance, consider the likelihood ratio $P(x_j|y=1) = \frac{\text{Num instances where } x_j=1, y=1}{\text{Num examples with } y=1}$, and suppose that there are few, or even no instances of $y = 1$. Adding a constant 1 to the numerator and 2 to the denominator can help prevent division by 0 and acts as a Bayesian

prior. Laplace smoothing introduces some bias, but it is hardly significant with a large number of training examples.

9.2.4 Multiclass Classification

The above examples assumed binary classification. For multiclass classification, there are two main options:

- Train a single classifier that can produce all classifications.
- Train a 1-vs-all binary classifier for all classes. This is more flexible, but often slower due to the number of separate classifiers that need to be trained, and introduces a class imbalance problem where the target class has relatively fewer points than the aggregation of all other classes.

9.3 Discriminative Learning

Discriminative learning attempts to model $p(y|x)$ directly. Logistic regression is a common example of discriminative learning.

Logistic regression is used to model binary outputs. Real target values are produced and are then mapped to a value between 0 and 1. Logistic regression assumes an exponential estimation parametric form of the distribution $P(Y|X)$ for the form $\sigma(w^T x) = \frac{1}{1+e^{-w^T x}}$. $\sigma(w^T x)$ is the probability that $y_i = 1$, and $1 - \sigma(w^T x)$ is the probability that $y_i = 0$. It follows that

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^n w_i X_i)}$$

$$P(Y = 0|X) = \frac{\exp(w_0 + \sum_{i=1}^n w_i X_i)}{1 + \exp(w_0 + \sum_{i=1}^n w_i X_i)}$$

9.3.1 Training

The goal of training is to find weights that maximize the likelihood $P(x_1, y_1, \dots, x_n, y_n|w) = \prod_{i=1}^n \sigma(w^T x_i)^{y_i} (1 - \sigma(w^T x_i))^{(1-y_i)}$, which is equivalent to minimizing $-\sum_{i=1}^n y_i \log(\sigma(w^T x_i)) + (1 - y_i) \log(1 - \sigma(w^T x_i))$

9.3.2 Decision Boundary

The decision boundary of logistic regression is the same as that of the naive Bayesian classifier (i.e. $\log \frac{P(y=1|x)}{P(y=0|x)} = 0$).

10 Support Vector Machines

10.1 Margins

A binary linear classifier can be defined as $h_{w,b}(x) = g(w^T x + b) = 1$ if $w^T x + b \geq 0$, otherwise -1 . A large margin involves points being far from the decision boundary. There are two main ways to measure margin.

10.1.1 Functional Margins

For a training example (x_i, y_i) , the functional margin is $\hat{\gamma}_i = y_i(w^T x_i + b)$. However, this is not a good measure of classification confidence, since w and b can be scaled to increase the margin without affecting the classification.

10.1.2 Geometric Margins

Geometric margins avoid the above problem with functional margins by normalizing the weight vector w . For a training example (w_i, y_i) , the geometric margin is $\gamma_i = y_i \frac{w^T x_i + b}{\|w\|}$.

10.1.3 Margin of Training Set

The margin of a training set is the smallest margin of its individual training examples. Minimizing the geometric margin for a binary classifier is equivalent to minimizing $\|w\|$.

10.2 Optimal Margin Classifier

The goal of an optimal margin classifier is to find the maximum geometric margin for a linearly separable data set. This can be denoted as solving $\max_{w,b} \frac{1}{2} \|w\|^2$ such that $y_i(w^T x_i + b) \geq 1$ for $i = 1, \dots, n$. This is the problem that a support vector machine (SVM) attempts to solve.

10.3 Constrained Optimization

10.3.1 Lagrange Multipliers

The above equation can be treated as a constrained optimization problem. These types of problems can be solved by Lagrange multipliers.

Given a constrained optimization problem of the form $\min_w f(w)$ such that $h_i(w) = 0$ for $i = 1, \dots, l$, the Lagrangian is defined as $L(w, \beta) = f(w) + \sum_{i=1}^l \beta_i h_i(w)$, where $\beta = \{\beta_1, \dots, \beta_l\}$

are called Lagrangian numbers. To solve for w and β we set the partial derivatives of L over w and β_i to 0.

Consider an example where we try to minimize $x + y$ such that $w^2 + y^2 = 1$. The Lagrangian is $L(x, y, \beta) = x + y + \beta(x^2 + y^2 + 1)$. Setting the gradient to 0 with respect to x , y , and β produces the system of equations $1 - 2\beta x = 0$, $1 - 2\beta y = 0$, and $x^2 + y^2 - 1 = 0$, which can be solved to yield $x = y = -\frac{1}{\sqrt{2}}$.

The generalized Lagrangian includes a $g_i(w) \leq 0$ constraint for $i = 1, \dots, k$. The generalized Lagrange is denoted as $L(w, a, b) = f(w) + \sum_{i=1}^k \alpha_i g_i(w) + \sum_{i=1}^l \beta_i h_i(w)$.

10.3.2 Primal and Dual Problem

Given the generalized Lagrangian above, we can define a primal $\theta_P(w)$ as $\max_{\alpha, \beta, \alpha_i \geq 0} L(w, \alpha, \beta)$. If f satisfies all constraints, then $\theta_P(w) = f(w)$, otherwise $\theta_P(w)$ is infinite. The “primal problem” consists of finding $p^* = \min_w \theta_P(w)$.

In the “dual problem”, we first minimize with respect to w , and then maximize with respect to α, β . The dual is defined as $\theta_D(\alpha, \beta) = \min_w L(w, \alpha, \beta)$, and the “dual problem” consists of finding $d^* = \max_{\alpha, \beta, \alpha_i \geq 0} \theta_D(\alpha, \beta)$.

The “max min” of a function is always less than or equal to the “min max”, so $d^* \leq p^*$. Under conditions known as Karush-Kuhn-Tucker (KKT) conditions, $d^* = p^*$.

10.4 Training

The high-level goal of training is to find w, b such that $\frac{1}{2}||w||^2$ is minimized and $y_i(w^T x_i + b) \geq 1, i = 1, 2, \dots, n$.

We can write the constraints for each training example i as $g_i(w) = 1 - y_i(w^T x_i + b) \leq 0$. The Lagrangian of the optimization problem is:

$$L(w, b, a) = \frac{1}{2}||w||^2 + \sum_{i=1}^n \alpha_i (1 - y_i(w^T x_i + b))$$

$\frac{\partial L(w, b, a)}{\partial w} = w - \sum_{i=1}^n \alpha_i y_i x_i$, which is set to 0 at $w = \sum_{i=1}^n \alpha_i y_i x_i$. Also, $\frac{\partial L(w, b, a)}{\partial b} = \sum_{i=1}^n \alpha_i y_i = 0$. Plugging these back into L , we end up with the dual problem for SVM being $\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j (x_i^T x_j)$ such that $\alpha_i \geq 0$ for $i = 1, \dots, n$ and $\sum_{i=1}^n \alpha_i y_i$ for $i = 1, \dots, n$.

If both f and g_i are convex, and all g_i can be satisfied for some w , then $d^* = p^* = L(w^*, b^*, \alpha^*)$, where w^* is the optimal weight vector set (primal solution) and α^* is the optimal support vector set (dual solution). In SVMs, both f and g_i are convex, and for linearly separable data, all g_i can be satisfied simultaneously. $g_i(w)$ that equal 0 are on the decision boundary, and are known as support vectors. The output classifier of a query point

x is computed as $h_{w,b}(\hat{x}) = \text{sign}(\sum_{i=1}^n \alpha_i y_i (x_i^T \hat{x}) + b)$. Since $\alpha_i > 0$ for only support vectors, only support vectors are considered in this computation.

10.5 Handling Non-Linearly Separable Data

Two options to model data that cannot be classified using a linear boundary are:

- Relaxing the constraints by allowing some points to be incorrectly classified by the margin. In this case, $\min_{w,b} \frac{1}{2} \|w\|^2$ such that $y_i(w^T x_i + b) \geq 1, i = 1, 2, \dots, n$ can be rewritten as $\min_w \frac{1}{2} \|w\|^2 + L(w^T x_i + b)$, where L is some loss function penalizing incorrect classifications or points close to the decision boundary. A common loss function is known as hinge loss, and is defined as $\max(1 - y_i(w^T x_i + b), 0)$.
- Allowing a non-linear decision boundary by finding a linear decision boundary of some mapping ϕ of features.

11 Kernel Methods

As mentioned in the section above, the dual form of SVM can be expressed as $\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j (x_i^T x_j)$ such that $\alpha_i \geq 0$ for $i = 1, \dots, n$ and $\sum_{i=1}^n \alpha_i y_i = 0$ for $i = 1, \dots, n$. If applying a feature mapping ϕ , the x_i^T and x_j terms become $\phi(x_i)^T$ and $\phi(x_j)$ respectively. It follows that classification of an input x is given by $\text{sign}(\sum_{i=1}^n \alpha_i y_i (\phi(x_i)^T \phi(\hat{x})) + b)$, where $\phi(x_i)^T \phi(\hat{x})$ is known as the kernel.

11.1 Motivation

A linear model does not make use of ϕ , which results in simple decision boundaries. Data that is not linearly separable is often separable in higher dimensions, introducing the need for more complex mappings ϕ .

The motivation for adding kernels is similar to that of expanding the feature space to include more terms. For example, consider a quadratic feature space, where a set of features x_1, x_2, \dots, x_m is mapped to $x_1 x_1, x_1 x_2, \dots, x_1 x_m, x_2 x_1, x_2 x_2, \dots, x_2 x_m, \dots$. Problems with such a large feature space are increased computational cost and an increased chance of overfitting. However, this feature space can be represented by the kernel function $K(x, z) = (x^T z)^2$, in which case its calculation takes $O(m)$ steps instead of $O(m^2)$. Kernels can be widely applied to reduce computational complexity by rewriting algorithms in SVM dual form and then transforming the dot product term into a kernel function.

11.2 Common Kernels

- **Radial basis function (RBF)/Gaussian Kernel** $K(x, z) = \exp(-\frac{\|x-z\|^2}{2\sigma^2})$: High when points are close, decreases as points are further apart.
- **Linear Kernel** $K(x, z) = x \cdot z$
- **Polynomial Kernel** $K(x, z) = (1 + x \cdot z)^d$: Feature expansion has monomial terms of power d .

11.3 Kernel Properties

Let a kernel matrix K for n points x_1, \dots, x_n be defined as $K_{ij} = K(x_i, x_j)$. A kernel K is considered to be a Mercer kernel if and only if for any finite set of n points, the corresponding kernel matrix is symmetric and positive semi-definite. Symmetric means that $K(x_i, x_j) = K(x_j, x_i)$. Positive semi-definite means that for all functions f that are square integrable (i.e. $\int f(x)^2 dx < \infty$) other than the 0 function, $\int \int f(x) K(x, z) f(z) dx dz > 0$.

This characterization is useful because it provides ways to construct kernels from other kernels. For example, adding, multiplying, or scaling two Mercer kernels always generate a Mercer kernel.

11.4 Kernel Perceptron Algorithm

By the perceptron representer theorem, the weight vector during a run of the perceptron algorithm can be described as a linear combination of the expanded training data (i.e. $w = \sum_{i=1}^n a_i \phi(x_i)$). We can alter the perceptron algorithm using this property to update a to be $w \cdot \phi(x) + b$ and w to be $w + y\phi(x)$.