Machine Learning

Kiar Fatah

February 25, 2020

Contents

If you can't explain something in
simple terms, you don't understand it.

Richard Feynman

1 Lecture 2

1.1 Decision Tree

A decision tree is a map of the possible outcomes of a series of related choices.

1.2 How to build a Tree

1. Choose the best question (according to the information gain), and split the input data into subsets.

2. Terminate: call branches with a unique class labels leaves (no need for further questions).

3. Grow: recursively extend other branches (with subsets bearing mixtures of labels).

1.3 How to ask the best questions?

The machine will automate the questions for the decision tree, however how does it obtain the best questions to ask? It does so by with the help of Gini impurity and information gain.

1.3.1 Gini impurity

Gini impurity is a measurement of uncertainty in a node. It is calculated by how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

$$1 - \sum_i p_i^2. \tag{1.1}$$

1.3.2 Entropy

Entropy is also a measurement of uncertainty in a node and can be applied instead of Gini impurity. It is calculated as

$$\sum_i -p_i log_2 p_i. \tag{1.2}$$

The information entropy is measured in bits, corresponding to the logarithmic base 2. Tossing a coin results in the entropy 1, hence it has 1 bit of information. For a real dice and a fake dice the value is 2.58 and 2.16 bits of information respectively.

1.3.3   Information gain

Information gain is how much the question reduces the uncertainty. The information is given by the first set of values for the internal node subtracted by the average value of the impurity of the split set.

1.3.4   Best question

The question with the ==highest value== on the information gain will be the one to ask.

1.4   Overfitting

Overfitting is when the learned models are overly specialised for the training samples. This occurs when the data is too noisy, not representative and when the model is too complex. This can be tackled by choosing a simpler model.

1.4.1   Pruning

The idea of reduced error pruning is to consider each node in the tree as a candidate for removal. A node is removed if the resulting pruned tree performs at least as well as the original tree over a separate validation dataset. The pruning is done on validation set, until it is harmful for the model.

1.4.2   Validation set

Due to pruning there is a need for validation dataset that is not training nor test data set. Therefore if there is access to a rich data set, it is separated to three categories. In essence the validation set is used to estimate prediction error for model selection (i.e. to determine hyperparameters).

2   Lecture 3

2.1   Introduction

How does one determine the right model, f, from the data? The simple concept for classification is by calculating the misclassification rate in consideration to the data. Assume the

following data is given $\mathcal{D} = (\boldsymbol{x_1}, y), ...$ is given. The misclassification rate is given by

$$err(f, \mathcal{D}) = \frac{1}{N} \sum Ind(f(\boldsymbol{x}) \neq y) \tag{2.1}$$

2.2   Curse of Dimensionality

1. Easy problems in low-dimensions are harder in high-dimensions, e.g training more complex models with limited sample data.

2. In high-dimensions everything is far from everything else, this is an issue in the nearest neighbour model, this causes issues with Nearest Neighbours.

3. Any method that attempts to produce locally varying functions in small isotropic neighbourhoods will run into problems in high dimensions.

2.3   The Bias-Variance Trade-off

Let us imagine we could repeat the modelling for many times – each time by gathering new set of training samples, D. The resulting models will have a range of predictions due to randomness in the underlying data set.

1. Variance: Refers to the amount by which $\hat{f}$, the prediction function, would change if we estimated it using a different training data set.

2. Bias: Refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.

This can be showcased in figure 1 In essence the relationship between complexity, bias and variance can be summarised as

1. Low model complexity implies high bias, low variance.

2. High model complexity implies low bias, high variance.

Note that bias and variance are equally important as we are always dealing with a single realisation of the data set.
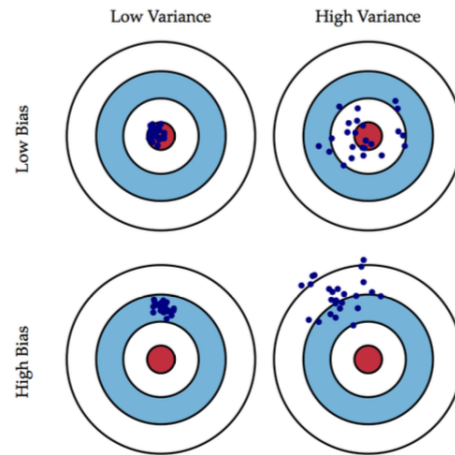
Figure 1: Variance versus Bias

3   Lecture 4

3.1   Linear Regression, A Parametric Method

Simple linear regression assumes theres a linear relationship between the output and input. Therefore a linear equation is assumed, parametric method. However, the coefficients are unknown and thus needs to be approximated.

The measurement of error is mean square. It is a convex function. Therefore the coefficients that minimises the sum of the squared error is given by the vector that sets the gradient of the sum of squared error to zero.

$$E_{in}(w) = ||Xw - Y||^2 \implies \frac{\partial E_{in}}{\partial w} = 2X^T(Xw - Y) = 0 \implies w = (X^TX)^{-1}X^TY \quad (3.1)$$

whereas w is the coefficients, X the data, Y the label and E_{in} the sum of squared errors.

3.2   RANdom SAmpling Consensus

Random sampling consensus or RANSAC is applied when the data set S contains outliers

1. Randomly select a (minimum number of) sample of s data points from S and instantiate the model from this subset.

2. Determine the set of data points S_i which are within a distance threshold t of the model. The set S_i is the consensus set of samples and defines the inliers of S.

3. If the subset of S_i is greater than some threshold T, re-estimate the model using all the points in Si and terminate

4. If the size of S_i is less than T, select a new subset and repeat the above.

5. After N trials the largest consensus set S_i is selected, and the model is re-estimated using all the points in the subset S_i

In essence RANSAC takes two random data points and fits a line. This is repeated until there is a good result of data points within the margin of the fitted line. RANSAC is applied when the dataset contains outliners.
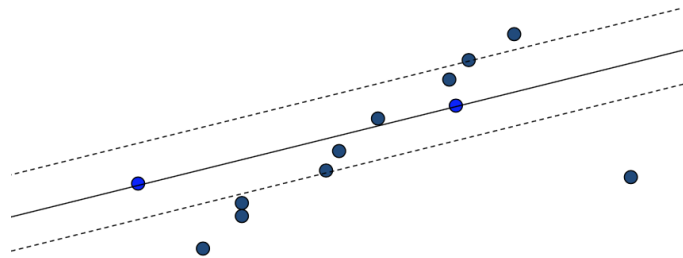
Figure 2: Line fitting between two data points.

3.3   Disadvantages with RANSAC

The threshold of the margin in the line is vulnerable.

1. If it is to high every fit is ranked equally.

2. If it is to low it leads to an unstable fit.

However, this issues are tackled within MLESAC that evaluates the quality of the consensus dataset by calculating its likelihood for better predictions.

3.4   k-NN Regression, A Non-parametric

The k-NN regression method is closely related to the k-NN classifier. Given a value for K and a prediction point x_0, k-NN regression first identiffies the K training observations that

are closest to x_0 represented by N_0. It then estimates $f(x_0)$ using the average of all the training responses in N_0. In other words

$$f(\hat{x_0}) = \frac{1}{K} \sum_{x_i \in N_0} y_i.$$

(3.2)

Note larger values of k provide a smoother and less variable fit (lower variance!)

3.5   Parametric or Non-parametric Methods?

1. If the parametric form is close to the true form of f, the parametric approach will outperform the non-parametric

2. As a general rule, parametric methods will tend to outperform non-parametric when there is a small number of observations per predictor (i.e. in a high dimension).

3. Interpretability stand point: Linear regression preferred to KNN if the test MSEs are similar or slightly lower.

3.6   Shrinkage Methods

A common way to fit models is by applying least square, however an alternative is it is possible to fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently that shrinks the coefficient estimates towards zero.

1. Among a large number of variables X in the model there are generally many that have little (or no) effect on Y

2. Leaving these variables in the model makes it harder to see the big picture, i.e. the effect of the "important variables"

3. Would be easier to interpret the model by removing unimportant variables (setting the coefficients to zero)

3.7   Ridge Regression

Similar to least squares but minimizes different quantity

$$RSS + \lambda \sum_i w_i^2$$

(3.3)

Note that the second term is called shrinkage penalty.

1. Shrinkage penalty is a small value when the coefficients w_i are close to zero.

2. The parameter λ controls the impact of the two terms. Hence the selection is critical.

3. For a higher value on λ the variance decreases and bias increases.

4. Ridge regression is harder to interpret.

3.8 The Lasso

Similar to least squares but minimizes different quantity

$$RSS + \lambda \sum_i |w_i| \tag{3.4}$$

4 Lecture 5

4.1 Axiomatic definition of probabilities

Given an sample space ω of all possible outcomes, and an event E, or set of outcomes from ω then

1. $P(E) \geq 0$ for all $E \subseteq \omega$.

2. If $E = \omega$ then the probability of ω is one. In other words the probability that one event will occur is equal to one.

3. if the events are a ocuntable sequence of pairwise disjoint events then the probability of the union of each event is equal to the probability of the summation of every event.

4.2 Random (Stochastic) Variables

A random variable is neither random nor a variable, it is a function that does exactly what we need, $X : \Omega \to \mathbb{R}$. The probability distribution function, so called pdf, of a random variable maps the range to positive numbers, $Pr(x) : X \to \mathbb{R}_{>0}$.

1. The probability of a random variables describes how the probability density is distributed over the range of X,

$$P[0 \leq X \leq 1] = \int_{x=0}^{1} Pr(x)dx. \tag{4.1}$$

2. X is distributed Pr(x) is written ore compactly as

$$X \sim Pr(x). \tag{4.2}$$

An example is

1. Ω - All possible graded exams of all students in class.

2. $X : \Omega \to \mathbb{R}$ - Random variable mapping exam outcome to score.

3. $X \sim Pr(x)$ - Gaussian distribution.

4.3   Types of Random Variables

1. Discrete random variable: Countable set

2. Continuous random variable: Uncountable set

4.4   Joint Probabilities

Consider two random variables X and Y. Observe multiple paired instances, some paired outcomes will occur frequently. This information is encoded in the joint probability density function $Pr(x, y)$.

4.5   Marginalization

The probability distribution function, PDF, of any single variable can be recovered from a joint distribution by summing for the discrete case and integrating for the continuous case.

$$Pr(x) = \sum_y Pr(x, y), (Discrete) \tag{4.3}$$

4.6   Conditional Probabilities

The conditional probability of X given that Y takes a value y gives

$$P(A|B) = \frac{P(A, B)}{P(B)}. \tag{4.4}$$

If it turns out the events are independent then

$$P(A|B) = \frac{P(A)P(B)}{P(B)}. \tag{4.5}$$

Bayes's theorem gives us

$$P(A|B) = P(A|B)P(B) = P(B|A)P(A) \implies P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{4.6}$$

4.7 Common Distributions

1. Bernoulli - is a discrete distribution that models binary trials.

2. Categorical - is a discrete distribution that determines the probability of observing one of k possible outcomes. Bernoulli distribution is a special case of categorical distribution.

3. Gaussian or Univariate normal distribution - is a continuous distribution.

4.8 Central Limit Theorem

The distribution of a linear combination of a large number of independent, identically distributed (iid) variables will tend to normal, regardless of the underlying distribution.

4.9 Expectation

Given a function, $f[*]$, that returns a value for each possible value x^* of the variable x and a probability $Pr(x = x^*)$ that each value of x occurs. The expected output of that fuction is calculated according to

$$\mathbb{E}[X] = \mu_X = \int x Pr(x) dx. \tag{4.7}$$

it is equal to the center of gravity of a distribution. However, for different choices of the function, $f[*]$, the interpretation is different. Hence the variance can be calculated as

$$Var[X] = \sigma_X^2 = \mathbb{E}[X - \mathbb{E}[X])^2], \tag{4.8}$$

and is interpreted as the spread of a distribution. The covariance is given by

$$\sigma_{X,Y} = \mathbb{E}[X - \mathbb{E}[X]](Y - \mathbb{E}[Y])], \tag{4.9}$$

it showcases how two variables vary together.

4.10   General Machine Learning Problem

if Y is discrete then it is classification, if it is continuous it is regression.

1. Learning: We estimate $Pr(\boldsymbol{x}, y)$ from data.

2. Inference: We estimate $Pr(y|\boldsymbol{X} = \boldsymbol{x})$ from data.

4.11   Bayes's Rule

1. $Pr(\boldsymbol{x}|Y = y)$: Likelihood represents the probability density of observing data $\boldsymbol{x}$ given the hypothesis Y = y

2. $Pr(Y = y)$: Prior represents the knowledge about Y before any observation.

3. $Pr(y|\boldsymbol{X} = \boldsymbol{x})$: Posterior represents the probability density of hypothesis y given observation $\boldsymbol{X} = \boldsymbol{x}$.

4. $Pr(\boldsymbol{X} = \boldsymbol{x})$: Evidence describes how well the model fits the evidence.

4.12   Probabilistic Regression

Regression via conditional probability:

1. Find the joint distribution of $\boldsymbol{X}$ and Y: $Pr(\boldsymbol{X}, y)$

2. Compute the posterior of Y: $Pr(y|\boldsymbol{X} = x) = \frac{Pr(\boldsymbol{x},y)}{Pr(\boldsymbol{X}=x)}$

3. Compute conditional expectation: $\mathbb{E}[Y|X = x]$.

Explicit regression model:

1. Define a deterministic model $y = f(\boldsymbol{x}) + \epsilon$.

2. Describe probability distribution of the error $\epsilon = y - f(\boldsymbol{x})$.

3. Estimate the parameters in $f(\boldsymbol{x})$.

4.13   Selecting the most probable hypothesis

Maximum A Posteriori (MAP) is about choosing hypothesis from Y with highest probability given observed data x:

$$y_{MAP}(\boldsymbol{x}) = argmax_{y \in \mathcal{Y}} Pr(\boldsymbol{X} = \boldsymbol{x}|Y = y)Pr(Y = y) \tag{4.10}$$

Maximum Likelihood (ML) is If we do not know prior distribution, then choose hypothesis with highest likelihood of generating the observed data:

$$y_{MLE} = argmax_{y \in \mathcal{Y}} Pr(\boldsymbol{X} = \boldsymbol{x}|Y = y) \tag{4.11}$$

5   Lecture 6

5.1   Introduction

Previous lecture Bayes' theorem was mentioned. However how does one go by to obtain the prior, likelihood and evidence from data?

5.2   Discriminative vs Generative Models

Discriminative modeling	Generative modeling
This models $Pr(y|\vec{x}, D)$ directly	This models $Pr(\vec{x}|y, D)$
Example: Logistic Regression	Naive Bayes

5.3   Parametric vs Non-parametric Inference

The posterior, $Pr(y|\vec{x}) = Pr(y|\vec{x}, \theta)$, distribution in consideration data is characterized by parameters θ.

Parametric Inference	Non-Parametric Inference
Estimate θ using data, D	Estimate $Pr(\theta|D)$
Computes $Pr(y|\vec{x}, \theta)$ to make inference	Compute $Pr(y|\vec{x}, D)$ from $Pr(y|\vec{x}, \theta, D)Pr(\theta, D)$
Learning corresponds to estimating θ	The number of parameters can grow with data
MAP & ML estimation	Bayesian methods

5.4   Maximum Likelihood (ML) Estimate

The ML estimation allows for approximation of posterior, $Pr(y|\vec{x}, \theta)$, and likelihood, $Pr(\vec{x}|y, \theta)$, by finding the parameter values that make the data most likely. The $\theta_M L$ is obtained through ML optimally which is defined as maximizing the likelihood of the data, D

$$\theta_{ML} = argmax_\theta P(D|\theta) = argmax_\theta \log P(D|\theta). \tag{5.1}$$

We can then approximate distributions given the data: $Pr(y|\vec{x}, \theta) \approx Pr(y|\vec{x}, \theta_{ML})$ and $Pr(\vec{x}|y, \theta) \approx Pr(\vec{x}|y, \theta_{ML})$. However note that a first source of confusion is ML parameter estimation is not ML regression/classification. The second one is Maximum a Posteriori (MAP) and Maximum Likelihood (ML) classification are different

$$y_{MAP} = argmax_y Pr(y|\vec{x}, \theta_{ML}), \tag{5.2}$$

$$y_{ML} = argmax_y Pr(\vec{x}|y, \theta_{ML}), \tag{5.3}$$

even with parameters θ estimated with the ML optimality criterion

$$y_{ML} = argmax_\theta Pr(D|y, \theta) = argmax_\theta \prod_n P(x_n|y_n).s \tag{5.4}$$

5.5   Curse of Dimensionality

1. Volume of feature space exponential in number of features

2. More features implies potential for better description of the objects but need more and more data to model Pr(x,y) well

However, for Naive Bayes classifier

1. All features (dimensions) regarded as conditionally independent.

2. Instead of modelling one D-dimensional distribution, model D one-dimensional distributions.

5.6   Naive Bayes Classifier

$\boldsymbol{x}$ is a vector $(x_1, ..., x_D)$ of attribute or feature values and Let $\mathcal{Y} = 1,2,...,$K be the set of possible classes. The MAP classification becomes

$$y_{MAP} = argmax_{y \in \mathcal{Y}} Pr(y|(x_1, ..., x_D)) = argmax_{y \in \mathcal{Y}} \frac{Pr((x_1, ..., x_D)|y)Pr(y)}{Pr((x_1, ..., x_D))} \tag{5.5}$$

,

$$argmax_{y\in\mathcal{Y}}\frac{Pr((x_1,...,x_D)|y)Pr(y)}{Pr((x_1,...,x_D))} = argmax_{y\in\mathcal{Y}}Pr((x_1,...,x_D)|y)Pr(y). \qquad (5.6)$$

The Naive Bayes assumption is

$$Pr((x_1,...,x_D)|y) = \prod_d Pr(x_d|y). \qquad (5.7)$$

The MAP classification with Naive Bayes becomes

$$y_{MAP} = argmax_{y\in\mathcal{Y}}Pr(y)\prod_d Pr(x_d|y). \qquad (5.8)$$

When to use Naive Bayes

1. Moderate or large training set available

2. Feature dimensions are conditionally independent given class (or at least reasonably independent, still works with a little dependence)

One issue with Naive Bayes is what if none of the training instances with target value y have attribute x_i? Then $Pr(x_i|y) = 0 \implies Pr(y)\prod_d Pr(x_d|y) = 0$. A simple solution to this is to add pseudocounts to all counts so that no count is zero. This is a form of regularization or smoothing.