Self Taught Machine Learning : Notes on Probability, Statistics, Linear Algebra, Statistical Learning, Theoretical(Practical) Machine Learning, Deep Learning Concepts

A. Chau

March 2023

1 You are good at something doesn't mean you really enjoy it

1.1 Probability

Random Variable(RV) : a variable that may happen in different values, depending on probability. E.g. The value of a fair dice ,possible values are 1,2,3,4,5,6 , each having probability of 1/6.

Some fundamental equations

Independence of RV(Discrete , Continuous) : $P(XandY) = P(X)P(Y)$

$$Var(X) = \sum_i (x_i - E(X))^2 P(X = x_i)$$

$$E(X) = \sum_i x_i P(X = x_i)$$

$$E(aX + b) = aE(X) + b$$

$$Var(X) = E(X^2) - (E(X))^2$$

Binominal Distribution : e.g. tossing a coin(fair or unfair) n times, number of times the head shows up

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, E(X) = np, Var(X) = npq$$

Poisson Distribution : (number of independent events happened AND N must be very large but no need infinite): Number of visits to a website per day.

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!},$$

Cumulative Distributive Function (CDF) : P(greater than x)

$$F(x; \lambda) = \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^i}{i!}$$

where λ is the mean

Central Limit Theorem (There is proof, not yet read it) : Whenever a random sample of size n is taken from **any** distribution with mean and variance, then the sample **mean** will be approximately normally distributed with population mean and variance. The larger the value of the sample size, the better the approximation to the normal. General rule of thumb is that result will be good if sample size n $>$ 30 , if skewed mean and variance, it requires more.
That is :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sample mean = population mean and normally distributed

$$Var(\bar{X}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

With CLT, if the sample data have sample means lie within 95% interval, we can claim that 95% the sample data comes from the population.

1.2 Hypothesis Testing with t-test,z-test,Z-test

H0(Null Hypothesis) - Assumption failed to be rejected
H1(Alternative Hypothesis) - Assumption rejected at some level significance
one-tail-test vs 2-tail-test, 95% confidence interval as example,
one tail $\rightarrow$ whether the value has increased/decreased. (Question : data fall in last max/min 5% of distribution)
2-tail $\rightarrow$ whether the value has changed. (Question : data fall in 2 ending 2.5 % section of distribution) p-value : The probability that the sample gives such a (or more extreme) result, given the condition that H0 is true.
It is tested against significant level (e.g. 95%/99%) for Hypothesis testing
p-value obtained by sample is used to compare against a critical value of a sig. level for H-Test
p-value not passed critical value, not sig. result , failed to reject H0
p-value passed critical value, sig. result , reject H0
e.g. sample = 100 , mean = 25 min , p-value = P(mean ¿ 25 min — H0 true)
Assume Sig. level = 95%

Type I Error : Incorrect rejection of a true H0, false positive
Type II Error : H1 is in fact true but we failed to reject H0, false negative

T-test : (sample size ¡ 30 , or (unknown population var AND sample size large) or known normal distribution)
1-sample-test : test whether mean of single population = target value (is mean this sample of this population = 30)
2-sample-test : test whether means of 2 population are different (does mean of height of female diff from mean of height of male)
Paired T : Test whether difference between 2 dependent = target value (does mean of weight diff after taking weight loss pill)

Z-test (assumed normal distribution , used when known population variance or sample size ≥ 30)
1-sample-test: test whether sample mean =/= population mean
2-sample-test: test wether means of 2 samples are different.
t-test use sample SD , z-test used population SD

1.3 Bayes Theorem, Likelihood, Maximum Likelihood Estimation(MLE)

The Bayes Theorem:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

θ : Parameter of data distribution, e.g. Means , Variances of bi(multi)-variate normal distribution, or estimator
X : sample data set
$P(\theta|X)$: Posterior
$P(X|\theta)$: Likelihood, given parameter θ), the probability that such a sample X will shows up
$P(\theta)$: Prior
$P(X)$: Evidence

Maximum Likelihood Estimation (MLE) : We want to infer θ, What parameter will give the maximum chance of seeing X ?

$$\theta_{MLE} = \arg\max_{\theta} P(X|\theta) = \arg\max_{\theta} \prod_i P(x_i|\theta)$$

If the function is convex , we can apply log likelihood, after taking log , if the fcn is increasing, it is still increasing. Hence, taking log does not change the maxima or minima

$$\theta_{MLE} = \arg\max_{\theta} \log P(X|\theta) = \arg\max_{\theta} \log \prod_i P(x_i|\theta) = \arg\max_{\theta} \sum_i \log P(x_i|\theta)$$

MAP (Maximum a posterier) is similar to MLE , but some weight is put to intial guess of θ 2 different ways of thinking, MLE : data speaks for itself(Machine Learning), MAP take into account the initial guess (Bayesian Statistics AAAAA)

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \propto P(X|\theta)P(\theta)$$

$$\theta_{MAP} = \arg\max_{\theta} P(X|\theta)P(\theta) \tag{1}$$

$$= \arg\max_{\theta} \log P(X|\theta) + \log P(\theta) \tag{2}$$

$$= \arg\max_{\theta} \log \prod_{i} P(x_i|\theta) + \log P(\theta) \tag{3}$$

$$= \arg\max_{\theta} \sum_{i} \log P(x_i|\theta) + \log P(\theta) \tag{4}$$

This means that in computing estimators , weight is put to prior, Special case : uniform distribution , MLE = MLP

1.4 Covariance Thing

In overvall - Positive : 2 RVs are positively related, Negative : 2 RVs are negatively related, Zero : unrelated

$$\mathrm{COV}[X, Y] = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{5}$$

$$\mathrm{Cov}(X, Y) = \mathrm{E}\left[(X - \mathrm{E}[X])\,(Y - \mathrm{E}[Y])\right] \tag{6}$$

$$= \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) \tag{7}$$

Covariance Matrix is the matrix with i,j element being , it is a positive semi-definite , symmetric matrix

$$\mathrm{COV}(x_i, x_j)$$

For a sample of vectors $x_i = (x_{i1}, \ldots, x_{ik})^\top, with, i = 1, \ldots, n$ the sample mean vector is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Q is Covariance Matrix :

$$Q = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$$

$$y^\top Qy = y^\top \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top\right) y$$

$$y^\top Qy = y^\top \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top\right) y$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left((x_i - \bar{x})^\top y\right)^2 \geq 0\,.$$

whish is SPD

Every positive semidefinite matrix only has eigenvalues ≥ 0
Remember Eigenvector and Eigenvalue

$$A\vec{v} = \lambda\vec{v}$$

$$\vec{v}^T A\vec{v} = \vec{v}^T(\lambda\vec{v}) = \vec{v}^T\vec{v}\lambda$$

Left is PSD, $\vec{v}^T\vec{v}$ is positive , then λ is also positive

1.5 Information , Shannon's Entropy

$$H(X) = -\sum_{x\in X} P_X(x)\log P_X(x). \tag{8}$$

The entropy measures the"AVEREGE amount of information"or "surprise" or "uncertainty" present in a random variable. (the number of questions to answer (the information to obtain) before we get the outcome)
How are these 3 concepts related ?
Consider Manchester United and other 15 Hong Kong secondary school soccer teams played against each other in a play off game for many years and we have to guess which which team was the champion. The chance of winning will probably be {0.9,0.01,0.01....} chance of winning , for this distribution , we have little suprise or uncertainty. If we can ask the question smartly , we can on AVERAGE ask fewer number of question to find the outcome (lets say if we choose to ask "Has MU won the final ?", we may often get the outcome in just 1 question)).

In the above case, we have little information. The distribution has little surprise and we high certiainty that MU will win the game
Mathematically and Intuitively uniform distribution has highest entropy.

In decision tree, we use Entropy formula to calculate max. Information gain , this the combination of greedy algorithm (to answer less question as possible and to make our answer as certain as possible),

1.6 Logistic Regression Reminder

we want to find θ where θ_i is the constant of x_i

$$h_\theta(x) = \sum_j \theta_j x_j = \theta^\top x$$

Our target is to minimize the Loss function , the MSE(Mean Squared Error)

$$J(\theta) = \frac{1}{2} \sum_i \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 = \frac{1}{2} \sum_i \left(\theta^\top x^{(i)} - y^{(i)} \right)^2$$

2 Glossory: Hill Walking

Generalization error: errors rate that a model worked on unseen data. i.e. test error. In ML, Stat.L, GE = measure of how accurate an algorithm to predict previously unseen data. sometimes = test error. It includes irreducible error (Noise), Bias : models' too simple(underfitting) ,Or Variance : model too complicated(overfitting), captured too many noise in training data.
Standard error : Standard error of mean.
Training Error : error (rate) obtained during training. e.g. misclassification rate during training
DEV error : error rate obtained during parameter tuning of the trained model.
Test Error : Error happening on validation validation.
Independent, Identically distributed(iid) : Independent – different samples have no correlations , formal P(x and y) = P(x)*P(y). Identically distributed – have the same distribution
Correlation vs Causation : Correlation does not means causation , sales of ice-cream has high correlation with opening hour of aircon, but there is a third factor (weather, temperature)

3 References:

https://bookdown.org/egarpor/inference/ – statistical inference
http://www.seas.ucla.edu/ vandenbe/ece133b.html
https://www.ise.ncsu.edu/fuzzy-neural/wp-content/uploads/sites/9/2018/11/Chapter9-Factorization-Full.pdf