

Statistical Inference

Equality of Opportunity - G... X

What do you want to learn?

Introduction

Why Google?

Module 1 Quiz

What it means to be AI first

Two stages of ML

ML in Google products

Demo: ML in Google products

Replacing heuristics

It's all about data

Framing an ML problem

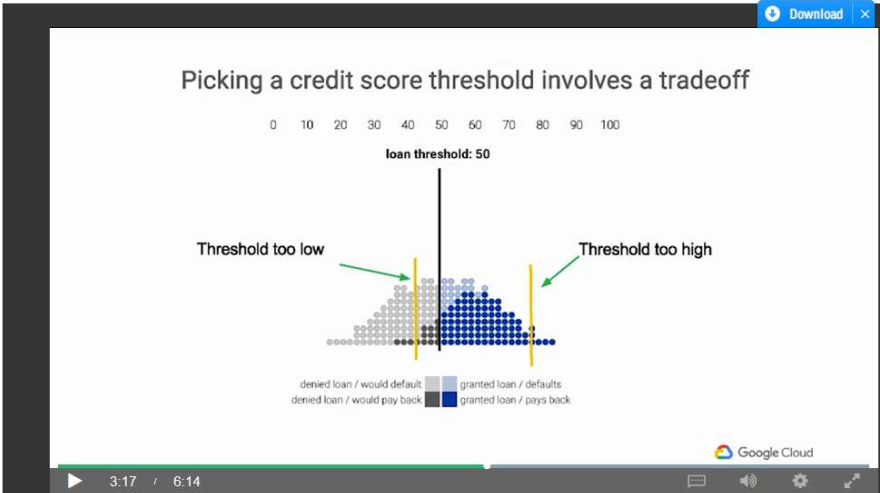
ML in applications

https://www.coursera.org/learn/google-machine-learning/lecture/06Kpw/equality-of-opportunity

Search...

ssahu

Picking a credit score threshold involves a tradeoff



Statistical Inference is the process of making conclusions about Population from Noisy Data (i.e., Sample) that was drawn from it.

⇒ $\text{Odd} = p / (1-p)$ => if p is the probability of occurrence

⇒ Variance, Covariance & Correlation

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

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \quad \begin{array}{l} \text{Normalized} \\ \text{Co-variance} \end{array}$$

⇒ PMF (mass function) for 'discrete' random variable... probability of all random values should add upto 1.0

⇒ PDF (density function) for 'continuous' random variable

⇒ Area(pdf curve between v_i & v_j) is total probability of variable taking values between v_i & v_j ... probability that a continuous variable can take one-value is ZERO as area of a line is ZERO

⇒ CDF.... Cumulative density function... $\Rightarrow F(x) = P(X \leq x)$

⇒ Survival function... $\Rightarrow S(x) = P(X > x)$

⇒ Cumulative prob of x taking a value ≤ 0.75 where PDF is of triangle shape with $0,0; 1,0; 1,2$, i.e., $p(x)=2x$

```
>> pbeta(0.75, 2, 1)
```

⇒ SURVIVAL gives probability of x taking a value > 0.75

```
>> qbeta(0.5, 2, 1)
```

=> 0.707... q=quantile function... beta=Bernaulli/binary dist

Conditional Probability / Bayes Theorem

⇒ Sensitivity = $P(+ | D)$

=> probability of tested +ve, given that subject has Disease

⇒ Specificity = $P(- | D_c)$

=> probability of tested -ve, given that subject didn't have Disease

⇒ Positive predictive value = $P(D | +)$

=> probability of having Disease, given than subject has tested +ve

⇒ Negative predictive value = $P(Dc \mid -)$

⇒ Prevalence of disease = $P(D)$

=> In the absence of any diagnostic testing

⇒ Diagnostic Likelihood Ratios (DLR+ve & DLR-ve)

⇒ The diagnostic likelihood ratio of a positive test, labeled DLR+, is

$$= P(+ \mid D) / P(+ \mid Dc)$$

$$= \text{sensitivity} / (1 - \text{specificity}).$$

⇒ The diagnostic likelihood ratio of a negative test, labeled DLR-, is

$$= P(- | D) / P(- | D^c)$$

$$= (1 - \text{sensitivity}) / \text{specificity}$$

Using Bayes rule, we have

$$P(D | +) = \frac{P(+ | D)P(D)}{P(+ | D)P(D) + P(+ | D^c)P(D^c)}$$

and

$$P(D^c | +) = \frac{P(+ | D^c)P(D^c)}{P(+ | D)P(D) + P(+ | D^c)P(D^c)}$$

Therefore, dividing these two equations we have:

$$\frac{P(D | +)}{P(D^c | +)} = \frac{P(+ | D)}{P(+ | D^c)} \times \frac{P(D)}{P(D^c)}$$

In other words, the post test odds of disease is the pretest odds of disease times the DLR_+ . Similarly, DLR_- relates the decrease in the odds of the disease after a negative test result to the odds of disease prior to the test.

So, the DLRs are the factors by which you multiply your pretest odds to get your post test odds. Thus, if a test has a DLR_+ of 6, regardless of the prevalence of disease, the post test odds is six times that of the pretest odds.

Expected Values

Population >> Sample >> Probability Mass/Density Function >> Characteristics of PMF/PDF are...

⇒ **Expected Values**

⇒ **Sample Quantiles**

Expected Values

⇒ Mean = centre of mass = Centre of distribution

⇒ Variance

⇒ $E(x) = \text{Mean} = \sum(x.p(x))$

⇒ $E(\text{dice}) = 1 \cdot 1/6 + 2 \cdot 1/6 + \dots + 6 \cdot 1/6 = 3.5$ ⇒ when fairly large number of trials are made, then we will get this average

⇒ ESTIMATOR IS UN-BIASED IF ITS DISTRIBUTIONS OF AVERAGE(s) WILL BE CENTERED AT/AROUND THE SAME VALUE WHERE THE POPULATION DISTRIBUTION

⇒ More precisely... take 10 dice... roll them... take SAMPLE VARIANCE of sides facing up on each dice... repeat over-&-over... plot variation of SAMPLE VARIANCE.... It will be centred around POPULATION VARIANCE

⇒ Take 20 dice... 30 dice.. or more

⇒ VARIANCE... how much the variable is varying around mean...

⇒ $\text{Var}(X) = E[(X - \text{mean})^2] = E[X^2] - (E[X])^2$ E is 'expected value'

⇒ For DICE

$$\text{Var}(X) = E[X^2] - (E[X])^2 = 15.167 - 3.5^2 = 2.917$$

$$E[X^2] = 1/6 \cdot 1^2 + 1/6 \cdot 2^2 + 1/6 \cdot 3^2 + 1/6 \cdot 4^2 + 1/6 \cdot 5^2 + 1/6 \cdot 6^2 = 15.167$$

$$E[X] = 3.5$$

⇒ For biased coin where probability of head(1) is p

$$E[X] = 0 \cdot (1-p) + 1 \cdot p = p$$

$$E[X^2] = 0^2 \cdot (1-p) + 1^2 \cdot p = p$$

$$E[X] = 0 \cdot (1-p) + 1 \cdot p = p$$

$$E[X^2] = E[X] = p$$

$$\text{Var}(X) = E[X^2] - E[X]^2 = p - p^2 = p(1-p)$$

$$\sigma^2 = \frac{\sum (x - \mu)^2}{n}$$

Population Variance =

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

Sample Variance =

⇒ $\text{Var}(X)$ of biased/unfair coin with head as '1' with probability 'p' (not 1/2) = $p \cdot (1-p)$

⇒ Expected Value of sample mean (this too is a random variable)

$$E[\bar{X}] = \mu$$

⇒ Variance of sample mean =

The variance of the sample mean is: $Var(\bar{X}) = \sigma^2/n$ where σ^2 is the variance of the population being sampled from.

Standard Error

Standard error, a.k.a. variability of Sample Mean, a.k.a., Regression Co-efficient = S/\sqrt{n} . **NOTE that "VARIANCE OF SAMPLE MEAN reaches ZERO for very large n", i.e., Sample Mean for large n is centred at Population Mean with ZERO VARIANCE**

- ⇒ STANDARD DEVIATION, a.k.a., STANDARD DEVIATION OF MEAN
- ⇒ **STANDARD ERROR, a.k.a., STANDARD ERROR OF MEAN**
- ⇒ The standard error indicates the likely **accuracy of the sample mean** as compared with the population mean.
- ⇒ The standard error decreases as the sample size (n) increases and approaches the size of the population.

- ⇒ Standard Error of Mean = $SEM = \frac{s}{\sqrt{n}}$ where 's' is Standard Deviation of Sample

S , the standard deviation, talks about how variable the population is

S/\sqrt{n} , the standard error, talks about how variable averages of random samples of size n from the population are

```
>> var(x)
>> sd(x)
```

Bernoulli's distribution of biased coin

Bernoulli's Distribution is special case of Binomial Distribution with $n=1$ (single coin)

$$P(X = x) = p^x (1 - p)^{1-x}$$

- ⇒ $P=1/2$ for fair, $x=0$ for tail & $x=1$ for head
- ⇒ $P(x=0) = 1/2$ for fair & p for unfair
- ⇒ $P(x=1) = 1/2$ for fair & $(1-p)$ for unfair

Mean and Variance

Let $X \sim \text{Ber}(p)$.

$$E(X) = 1 \times p + 0 \times (1 - p) = p$$

$$E(X^2) = 1^2 \times p + 0^2 \times (1 - p) = p$$

$$V(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1 - p)$$

Binomial distribution of multiple biased coins

A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment. A sequence of outcomes is called Bernoulli process.

```
>> choose(n,x)          ==> nCx
>> pbinom(...)          ==> probability of binomial distribution
>> pbinom(6, size=8, prob=0.5, lower.tail=FALSE) ==> 50% for boy/girl... prob of 7 or more girls out of 8 births
```

NORMAL & STANDARD NORMAL distribution

If $X \sim N(\mu, \sigma^2)$ then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

If Z is standard normal

$$X = \mu + \sigma Z \sim N(\mu, \sigma^2)$$

```
>> qnorm(0.95, mean=0, sd=1.0)      ## 95 %ile of N(mu=0, sigma=1)

>> pnorm(1160, mean=1020, sd=50, lower.tail=FALSE) ## prob of getting more than 1160 clicks in a day
>> pnorm(2.8, lower.tail=FALSE)      ## 1160-1020/50 = 2.8 to convert Normal to Standard Normal (mean=0, sd=1)

>> qnorm(0.75, mean=1020, sd=50)    ## what no of daily ad clicks would represent 75% of days have fewer clicks
```

Poisson Distribution

Lambda is average frequency; Its key feature is that mean == variance

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

Poisson random variables are used to model rates
 $X \sim \text{Poisson}(\lambda t)$ where
 - $\lambda = E[X/t]$ is the expected count per unit of time
 - t is the total monitoring time

- ⇒ Poisson dist with mean 2.5 people show up per hour at bus stop; if one waits for 4 hours, what is the prob of 3 or fewer people
 >> ppois(3, lambda=2.5*4)
- ⇒ Poisson approx. to Binom ($np < 10$, $n \geq 20$, $p \leq 0.05$... where $\lambda = n \cdot p$)
 >> pbinom(2, size=500, prob=0.01) => coin success prob is 0.01, flip 500 times; find prob of 2 or fewer success
 >> ppois(2, lambda=500*0.01) => same above use case tried through Poisson-approximating-binom

CENTRAL LIMIT THEOREM... Asymptotics

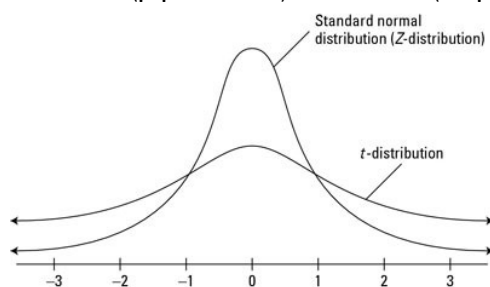
Distribution of averages is often normal, even if the distribution that the data is being sampled from is very non-normal!

Limits of random vars

- ⇒ **Law of large numbers (LLN)** – sample mean limits to population mean
- ⇒ **Central Limit Theorem (CLT)** – Distribution of average is often normal $N(\mu, \sigma^2/n)$ as the sample size increases (absolute value of confidence interval narrows inversely with 'n'), irrespective of sample distribution being non-normal.... NOTE that sample std dev (a.k.a., Standard Error) = population std dev/n
 - Standard Error –vs- Confidence Interval.... Here we are trying to find confidence interval between mean would be found... and NOTE that “distribution of average/mean is often normal....”. Best reference - <http://www.stat.wmich.edu/s160/book/node46.html>
 \bar{X} , is approximately normal with mean μ and sd σ/\sqrt{n}
 probability \bar{X} is bigger than $\mu + 2\sigma/\sqrt{n}$ or smaller than $\mu - 2\sigma/\sqrt{n}$ is 5%
 - $\bar{X} \pm 2\sigma/\sqrt{n}$ is called a 95% interval for μ
 - low & high tails 2.5% = 100 – 97.5; qnorm(0.975) = 1.96 ~ 2
 >> mean(x) + c(-1, 1)*qnorm(0.975)*sd(x)/sqrt(length(x))
 - 100 voters; 56 expected to vote in favour; find confidence interval for 95% win => voters are unbiased coin with p=0.56
 >> 0.56 + c(-1,1)*qnorm(0.975)*sqrt(0.56*0.44/100)
 >> binom.test(56, 100)\$conf.int ## same result as above as its default setting is for 95% confidence

Confidence Interval

Z-distribution (population data) & T-distribution (sample data, not population) are both different form of Normal Distribution



- ⇒ T-dist have heavier tails; mean=0 (always) => it is indexed by just 1 param (degree of freedom)
- ⇒ T-distribution is useful for estimating Population Mean from Sample Mean for small-n & SD-unknown
- ⇒ DegreeOfFreedom-of-T-dist = n-1
- ⇒ Unlike other normal dist where estimates are dependent on 2 params (mean & sd), estimates of T-dist is dependent on just DOF
- ⇒ T-quantiles are more conservative (wider) than Z-quantiles
- ⇒ As 'n' increases, T becomes Z
- ⇒ Should not be used for skewed or binary data

T-interval

Assumes that data are IID Normal, though it is robust to this assumption => hence, if data distribution is just symmetric & mound shape, then T-interval can be safely used for Confidence Interval estimation

- ⇒ Paired Observations can be analysed using T-intervals on their differences
- ⇒ Not suitable for SKEWED data distribution, as it does not make sense to centre the interval at mean
- ⇒ Not suitable for highly DISCRETE or BINARY distribution, i.e., other intervals are available

- ⇒ $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ follows Gosset's T-Distribution with n-1 DOF, i.e., this does not follow Gaussian distribution => but if Sigma (population variance) is used instead of S (sample variance), then that distribution is Gaussian

⇒ T Confidence Interval $\bar{X} \pm t_{n-1} S / \sqrt{n}$, where t_{n-1} is T-Quantile with n-1 DOF

Paired t-Confidence Interval

- ⇒ Sample is split in subsequent groups & subjects matched
- ⇒ [??] Why is sample split into groups?


```
>> data(sleep)
>> g1 <- sleep$extra[1:10]
>> g2 <- sleep$extra[11:20]
>> difference <- g2 - g1
>> mn <- mean(difference)
>> s <- sd(difference)
>> n <- 10
```
- ⇒ All t.test methods give same T-Quantile confidence interval


```
>> mn + c(-1,1) * qt(0.975, n-1) * s / sqrt(n)
>> t.test(difference)
>> t.test(g2, g1, paired=TRUE) // PAIRED T.TEST
>> t.test(extra ~ 1(relevel(group, 2)), paired = TRUE, data = sleep)
```

Independent group t-confidence intervals, i.e., Confidence interval for the difference of two population means

No pairing as there is no matching of the subjects. GOAL is to COMPARE the MEAN blood pressure of 2 groups (treatment –vs- placebo, but subjects are different in both groups)

- ⇒ <http://www.kean.edu/~fosborne/bstat/06b2means.html>
- ⇒ where the population variances are known (use z-interval)
- ⇒ where the population variances are unknown but equal (use t-interval)
- ⇒ where the population variances are unknown but unequal (use t')
- ⇒ Assumption.... CONSTANT POPULATION VARIANCE (though unknown) ACROSS BOTH GROUPS => reasonable assumption due to randomization => S_x & S_y may be slightly different (mostly due to different n_x & n_y)
- ⇒ Independent t-conf interval.... S_p is pooled variance

$$\bar{Y} - \bar{X} \pm t_{n_x + n_y - 2, 1 - \alpha/2} S_p \left(\frac{1}{n_x} + \frac{1}{n_y} \right)^{1/2} \Rightarrow \text{DOF is } n_x + n_y - 2; \text{ last part is std-error } (s / \text{sqrt}_n)$$

$$S_p^2 = \{(n_x - 1)S_x^2 + (n_y - 1)S_y^2\} / (n_x + n_y - 2)$$

- ⇒ Comparing SBP for 8 contraceptive users –vs- 21 controls


```
>> data(sleep)
>> x1 <- sleep$extra[sleep$group == 1]
>> x2 <- sleep$extra[sleep$group == 2]
>> n1 <- length(x1)
>> n2 <- length(x2)
>> sp <- sqrt(((n1 - 1) * sd(x1)^2 + (n2 - 1) * sd(x2)^2) / (n1 + n2 - 2))
>> md <- mean(x1) - mean(x2)
>> semd <- sp * sqrt(1/n1 + 1/n2)
>> md + c(-1, 1) * qt(0.975, n1 + n2 - 2) * semd
```

-or-

>> t.test(x1, x2, paired = FALSE, **var.equal = TRUE**)\$con

Un-equal Population Variance (though unknown), but then this kind of sample data does NOT have t-distribution, hence t'-kind-of-approximate distribution is assumed and confidence interval is calculated as per below

- ⇒ can use R command >> t.test(..., var.equal=FALSE)

$$\bar{Y} - \bar{X} \pm t_{df} \left(\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y} \right)^{1/2}$$

$$df = \frac{\left(S_x^2 / n_x + S_y^2 / n_y \right)^2}{\left(\frac{S_x^2}{n_x} \right)^2 / (n_x - 1) + \left(\frac{S_y^2}{n_y} \right)^2 / (n_y - 1)}$$

Hypothesis Testing

- ⇒ Statistical Inference are of type either ESTIMATION or HYPOTHESIS TESTING
- ⇒ [Best Article] <http://medianetlab.ee.ucla.edu/papers/HypothesisTesting.pdf>

A reasonable strategy would reject the null hypothesis if \bar{X} was larger than some constant, C

Typically, C is chosen so that the probability of a Type I error, α , is .05 (or some other relevant constant)

α = Type I error rate = Probability of rejecting the null hypothesis when, in fact, the hypothesis is correct

Standard error of the mean $10/\sqrt{100} = 1$

Under H_0 $\bar{X} \sim N(30, 1)$

We want to choose C so that the $P(\bar{X} > C; H_0)$ is 5%

The 95th percentile of a normal distribution is 1.645 standard deviations from the mean

If $C = 30 + 1 \times 1.645 = 31.645$

• If $C = 30 + 1 \times 1.645 = 31.645$

- Then the probability that a $N(30, 1)$ is larger than it is 5%

- So the rule "Reject H_0 when $\bar{X} \geq 31.645$ " has the property that the probability of rejecting H_0 is 5% when H_0 is true (for the μ_0 , σ and n given)

⇒ Reject NULL Hypothesis if $\sqrt{n}(\bar{X} - \mu_0)/s > Z_{1-\alpha}$, i.e., when $X_{\text{mean}} > C$ (this formulae is for data with Normal distribution – similarly use T-Quantile for T-distribution)

⇒ 2-sided Hypothesis Testing usually does not make scientific/intuitive sense, but it is still used

⇒ Null Hypothesis = H_0 (status-quo), H_a (alternative hypothesis)

TRUTH	DECIDE	RESULT
H_0	H_0	Correctly accept null
H_0	H_a	Type I error
H_a	H_a	Correctly reject null
H_a	H_0	Type II error

⇒ Type-I Error Rate (α) = Probability of REJECTING H_0 when in fact the H_0 is correct/true

⇒ Type-II Error Rate (β) = Probability of FAILING TO REJECT H_0 when in fact the H_0 is incorrect/false

⇒ α & β are inversely proportional

⇒ Overall goal is to maintain a balance between low- α (0.05 or 5% upper tail) & high-Power (i.e., low- β)

Hypothesis Testing >> Power Analysis

⇒ Power – Probability of REJECTING Null Hypothesis when it is FALSE

⇒ Power = 1 – Beta (where Beta is probability of FAILING TO REJECT Null Hypothesis when it is FALSE)

⇒ Power is more with bigger & bigger 'n', useful at design time

⇒ Calculating POWER for Gaussian Data

We reject if $\frac{\bar{X} - 30}{\sigma/\sqrt{n}} > z_{1-\alpha}$

- Equivalently if $\bar{X} > 30 + Z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$

Under H_0 : $\bar{X} \sim N(\mu_0, \sigma^2/n)$

Under H_a : $\bar{X} \sim N(\mu_a, \sigma^2/n)$

```
z <- qnorm(1 - alpha)
pnorm(mu0 + z * sigma/sqrt(n), mean = mua, sd = sigma/sqrt(n), lower.tail = FALSE)
```