BTRY 4090: Spring 2009
Theory of Statistics

Guozhang Wang

September 25, 2010

1 Review of Probability

We begin with a real example of using probability to solve computationally intensive (or infeasible) problems.

1.1 The Method of Random Projections

1.1.1 Motivation

In information retrieval, documents(or images) are represented as vectors and the whole repository is represented as a matrix. Some similarity, distance and norm measurements between documents(or images) involve matrix computation. The challenge is the matrix may be too large to store, and compute.

The idea is to reduce the matrix size while at the same time preserves characteristics such as Euclidean distance, inner products between any two rows.

1.1.2 Random Project Matrix

Replace original matrix A ($\in \mathbb{R}^{D\times n}$) by B ($\in \mathbb{R}^{n\times k}$) = A $\times$ R ($\in \mathbb{R}^{D\times k}$), where k is very small compared to n and D, and each entry in R is i.i.d sampled from $N(0, 1)$. At the same time, $E(BB^T) = AA^T$.

The probability problems involved are: the distribution of each entry in R; distribution of the norm for each row in R; the distribution of the Euclidean distance for each row in R; the error probabilities as a function of k and n.

1.1.3 Distribution of Entries in R

Since the entries of R are from normal distribution, its linear combination is also normal distributed with 0 mean and $\sum u_{j,i}^2$.

1.1.4 Distribution of Euclidean Norm

From the computational formula, we know that $Var(X) = E(X^2) - (E(X))^2$, we can get an unbiased estimator of the Euclidean norm $\hat{m_1} = \frac{1}{k}\sum_{j=1}^{k} |v_{i,j}|^2$. Since $v_{i,j}$ is i.i.d., $\hat{m_1} \times k$ has a Chi-squared distribution with k degrees of freedom.

And since we know that the mean for Chi-squared distribution is k, and variance is $2k$, we can get the variance of $\hat{m_1} = \frac{2 \times m_1^2}{k}$, where m_1 is the true value.

The coefficient of variation $\frac{Var(\hat{m_1})}{m_1^2} = \frac{2}{k}$, which is independent of m_1. This indicate that this is a good estimator with low relative variation. One note is that coefficient of variation has the assumption that the variation would increase as the real value itself increases.

1.1.5 Distribution of Euclidean Distance

Has the similar result as the Euclidean Norm.

1.1.6 Estimation on Inner Product

The estimator of inner product $\frac{1}{k}\sum_{j=1}^{k} v_{i,j} v_{k,j}$ is unbiased; however, the variance is $\frac{m_1 m_2 + a^2}{k}$, and thus the coefficient is not independent of a. One simple illustration is that when two vectors are almost orthogonal (which is common in high dimension space, where two vectors are orthogonal with probability close to 1), a is close to 0, but coefficient of variation is close to infinity. Therefore random projections may not be good for estimating inner products.

One note here is that this problem is due to the random sampling with entries of R, which is typical and hard to resolve.

1.1.7 Summary

This elegant method is suitable for approximating Euclidean distances in massive, dense and heavy-tailed (some entries in certain rows are excessively large) data matrix; However, it does not take advantage of the data sparsity. Another note is that it has intrinsic relationship with SVM (which is aimed at solution sparsity, but not data sparsity; methods like PCA takes advantage of data sparsity).

In real applications, the random matrix R can be applied only once. Since even we have multiple iterations of reduction and take the average value of estimators, the variance is the same.

1.2 Capture and Recapture Methods

1.2.1 Motivation

Consider in the Database query cost estimating process, the order of the join operator is crucial to the query performance, and is dependent on the estimate on the size of intermediate results. The size of intermediate results can not be known exactly before the join is operated. However, by sampling the tuples and operate the "mini-join" the sizes can be estimated using capturing and recapturing (sampling and mini join) methods.

Note this method has several important assumptions: 1) the total population does not change between capture and recapture; 2) the recapture process is random.

1.2.2 Estimation using Sampling

The method has the following steps:

1. Use combination counting rules to compute the probability of the recapture event.
2. After the probability is formalized as the function of the target population, compute maximum likelihood population.
3. The maximum likelihood value can be computed by observing the $ratioofsuccessiveterms$. Another way is plotting the curve and find the peak value (log form is suggested since the exponent arithmetic may be "explosive").

1.3 Bivariate Normal Distribution

A good property of normal distribution is that if the joint distribution is normal, then the marginal and conditional distribution is also normal. Furthermore, the linear combination of normals is also normal.

1.3.1 Bivariate Normal to Random Projection

Random projection utilizes this property to compute the variance of the unbiased estimators. Note here the variance of the estimator (which can also be treated as a random variable) is not the same as the variance of the population. The key idea is:

$$E(v_1, v_2)^2 = E(E(v_1^2, v_2^2|v_2)) = E(v_2^2 \times E(v_1^2|v_2))$$

Note v_2 is treated as a constant when it is the dependent variable of the conditional distribution, and $E(v_1^2|v_2)$ can be computed from $E(v_1|v_2)$ and $Var(v_1|v_2)$.

1.3.2 Moment Generating Function (MGF) to Random Projection

MGF also can be utilized to simplify the computation of the estimators for random projection. The basic procedure has two steps:

1. Use some known MGF to derive the estimator's MGF (for example, the function for normal distribution is $exp(\mu t + \theta^2 t^2/2)$, for chi-square distribution is $(1-2t)^(-\frac{k}{2}))$
2. Use the MGF to get the n^{th} moment of the estimator: $M_X^{(n)} = E[X^n e^{tX}]$.

One note is that when the first moment (mean) of the estimator a is α, the n^{th} moment of $a - \alpha$ is 0 only if the distribution is symmetric.

1.4 Tail Probabilities

The tail probability $P(X > t)$ or $P(|\bar{X} - X| \geq \epsilon X)$ is extremely important, since it tells what is the probability that the error between the estimated value and the true value exceeds an ϵ fraction of the true value. Thus by studying *how fast* the error decreases, it can also imply the sample size to achieve some required accuracy. One note is that if the *event* is the same, then the probability is the same.

On the other hand, it often requires numerical methods to exactly evaluate $P(X > t)$, and therefore one can give the tail probability upper bounds instead of the exact probability.

1.4.1 Tail Probability Inequalities

Before we give several tail probability inequality theorems, we would like to note that each theorem have some assumptions which limit its applicability (eg, Markov's Inequality assumes that the variable is non-negative and the first moment exists).

Theorem 1.1 (Markov's Inequality) *If X is a random variable with $P(X \geq 0) = 1$, and for which $E(X)$ exists, then:*

$$P(X \geq t) \leq \frac{E(X)}{t}$$

Markov's Inequality only uses the first moment and hence it is not very accurate.

Theorem 1.2 (Chebyshev's Inequality) *Let X be random variable with mean μ and variance σ^2. Then for any $t > 0$:*

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

Chebyshev's Inequality only uses the second moment. One note is that making error depend on the variance may be more reasonable than making it depend on the mean (eg, if the mean is 0, then Markov's Inequality is useless).

Theorem 1.3 (Chernoff's Inequality) *Let X be a random variable with finite MGF $M_X(t)$, then for any ϵ:*

$$P(X \geq \epsilon) \leq e^{-t\epsilon} M_X(t), \text{ for all } t > 0$$
$$P(X \leq \epsilon) \geq e^{-t\epsilon} M_X(t), \text{ for all } t < 0$$

The advantage of Chernoff's Inequality is that by choosing different t, one can get a family of bounds on the distribution. Then by choosing t to minimize the upper bound, this usually leads to accurate probability bounds, which decrease *exponentially fast*.

1.4.2 Sample Size Selection Using Tail Bounds

Since the variance of the estimator usually decrease with the number of samples (eg, the estimator of the mean of normal distribution is also normally distributed with variance $\frac{\sigma^2}{k}$). Thus by inputting this variance into the above inequality theorem, we can get the appropriate sample size to satisfy:

$$P(|\bar{X} - \mu| \geq \epsilon\mu) \leq \delta$$

for any δ. k is affected by

- δ: level of significance, lower value causes larger k.
- $\frac{\sigma^2}{\mu^2}$: noise/singal ratio, higher value causes larger k.
- ϵ: accuracy, lower value causes larger k.

2 Limit Theorem

2.1 The Law of Large Numbers

From the CLT, we can approximately get the *rate of convergence* of the LLN. Since $\bar{X} \approx N(\mu, \sigma^2/n)$, we have $Var(\bar{X}) = E((\bar{X} - \mu)^2) \approx \frac{\sigma^2}{n}$. Now we want to measure the error of the estimator by computing $E(|\bar{X} - \mu|)$ (Note we do not use square here to have the same scale for estimator μ). And from $Var(\bar{X})$ we get this expected value to be $\frac{\sigma}{\sqrt{n}}$. Therefore the rate of the convergence of LLN is $\frac{1}{\sqrt{n}}$.

Note this rate of convergence is a *worst case* distribution, actual rate of convergence depends on how similar the distribution is to the normal distribution.

The rate of convergence of *Monte Carlo method* is dependent on how many samples used to compute the average (in other words how many intervals are generated by dividing the space), and thus is $\frac{1}{n}$. Thus Monte Carlo has a faster rate of convergence, although in high dimension this difference becomes smaller.

2.2 Central Limit Theorem

There are two different theorems for convergence in distribution: *Continuity Theorem* is different from *Central Limit Theorem*. Continuity theorem tell us for some distributions (eg, Poisson, Gamma, etc), when the parameters of the distribution approaches some limit value, they would approaches to Normal Distribution. On the other hand, central limit theorem tell us the average of a sequence of i.i.d samples from *arbitrary* distribution approaches Normal distribution as the length of the sequence approaches to infinity. It does not have the requirement on the parameter limiting property of the distribution but limit the result only to the average value. Both of these theorems can be derived from MGF

Non-rigorously, we may say $\bar{X}$ is approximately $N(\mu, \frac{\sigma^2}{n})$.

The fitness of the Normal distribution depends on 1) whether the distribution is near symmetric, and 2) whether this distribution has heavy tails. The tail probability discuss before tell us the heavier the tail is, the more sample are needed to illustration the approximation.

3 Survey Sampling

From here we start the journey of statistics. The difference between probability and statistics is that probability is more like mathematics and statistics is more like science that applies maths. One important application of statistics is to obtain information about a large population by examining only a small fraction (called samples) of that population. It derives estimator of the characteristics (eg, mean, variance) of the population through the sample characteristics. The functions taking the samples as input to output the estimator can be treated as "statistics".

3.1 Simple Random Sampling

For simple random sampling (without replacement):

- Sample mean is the unbiased estimator of population mean, and also can be derived to population total.

- The accuracy of the sample mean depends on the sample mean's variance.
- Sample mean's variance is dependent on the population variance, sample size, and *finite population correction factor* due to realistic non-replacement sampling.
- Population variance can be unbiased estimated using sample variance, total size and sample size.
- Therefore the accuracy of sample mean as estimator of population mean depends on sample variance in probability.

One note is that for non-replacement sampling, when the sample size n approaches the population size N, the variance of the estimator $\frac{\sigma^2}{n}(\frac{N-n}{N-1})$ becomes 0. On the other hand, for replacement sampling, even when the sample size is the population size, the variance of the estimator is $\frac{\sigma^2}{N}$ but not 0. This is because for non-replacement sampling, if all the population is drawn the sampling result is always the same, therefore the estimator becomes unchanged; for replacement sampling, even if sampling size is N, some elements might be sampled multiple times while some others never sampled, which still causes the variance.

Is it always beneficial to remove the bias? It depends on the tradeoff between bias and variance (remember the $MSE = bias + variance$). If $E[\hat{\sigma}] = AE[\sigma^2]$, we can always get the unbiased estimator $\frac{\hat{\sigma}}{A}$. But if A is smaller than 1 then the variance $Var[\frac{\hat{\sigma}}{A}] = \frac{1}{A^2}Var[\hat{\sigma}]$ will be increased. If the increase of the variance overwhelm the decrease of the bias, it is not suggested to remove the bias. Furthermore, the choice of the sample size depends on the overall error (usually measured by MSE) which requires considering both bias (noted as noise/signal ratio) and variance.

4 Parameter Estimation

The task of parameter estimation is to estimate parameters $\theta_1, ..., \theta_k$ of the unknown density function (or called model) from n samples $X_1, ..., X_n$ generated from the density function.

Since from CLT, we can approximate the density function using the normal distribution. This gives us the density function "family" and ease our task to only estimate parameters μ and σ. But there are also conditions where even the density function family is not known and we have to 1) search in the model space; 2) estimate parameters.

4.1 Three Estimation Methods

In general, there are three estimation methods to estimate parameters:

- The method of moments: generate k equations using the first k moments to deal with the k unknowns.
- The method of maximum likelihood.
- The Bayesian method.

4.2 Method of Moments

Method of moments only applies to small number of parameters and closed form of the parameter/moment equations. That is mainly due to the computation issues: some high order moments of samples are computational impossible to have. Besides, the moment estimators are usually biased, although they are *consistent*. Note that consistency regards to properties when n approaches infinity, while bias and variance regards to properties when n is fixed.

We choose the lowest moments possible since as the order of the moments increases, the accuracy of the estimator decreases, since the square/cubic/... will enlarge the variance. For those densities that are not familiar, we have to first construct the equations by integrating the cdf.

4.2.1 Method of Maximum Likelihood

One note is that if after derivation the equations are nonlinear or even more complicated, an iterative scheme (like binary search) is needed to find the ML value of the parameters. Another numerical iterative method is "Newton Method" which utilize the taylor expansion to iteratively find points that is closer and closer to the maximum/minimum point. Actually, one-step-newton-method starting at a good point already works well. More iterations do not help much. Further, sometimes people can use the moment estimator as the starting point to use newton method.

4.2.2 Large Sample Theory for MLE

Assume i.i.d samples of size n, $X_i, i = 1$ to n, with density $f(x|\theta)$. Large sample theory says as $n \to \infty$, MLE estimator is asymptotically unbiased and normal, with mean θ and variance $\frac{1}{nI(\theta)}$, approximately. $I(\theta)$ is the *Fisher Information* of θ. For normal and binomial distribution, this "asymptotic" variance of MLE is also exact.

This "asymptotic" variance is also applied to computing the "asymptotic" *efficiency* between the unbiased estimators.

Definition 4.1 *Given two unbiased estimates, $\hat{\theta}_1$ and $\hat{\theta}_2$, the efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$ is*

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}$$

Given two asymptotically unbiased estimates, $\hat{\theta}_1$ and $\hat{\theta}_2$, the asymptotic relative efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$ is computed using their asymptotic variances (as sample size goes to infinity).

Theorem 4.2 (Cramér-Rap Inequality) *Any unbiased estimator T of θ where $f(x;\theta)$ is the density function where n samples are drawn. Then under smoothness assumption $f(x;\theta)$:*

$$Var(T) \geq \frac{1}{nI(\theta)}$$

Thus under reasonable assumptions, MLE is optimal or asymptotically optimal among all unbiased estimators.

4.2.3 Sufficiency

A statistic $T = T(X_1, X_2, ..., X_n)$ of i.i.d samples $X_1, X_2, ..., X_n$ from density $f(x;\theta)$ is said to be sufficient for θ if we can gain no more knowledge about θ even we see the whole samples besides the statistic (which mean, the conditional distribution given T does not depend on θ).

A necessary and sufficient condition for T to be sufficient for a parameter θ is that the joint probability density (mass) function factors.

The advantage of sufficiency is that, a sufficient statistic is "sufficient" to infer the density function parameter, and we do not need to keep the samples any more once we have the statistic, so storing only one value is much more space efficient compared with storing the whole set of sample values.

The Cramér-Rao Inequality tells us the MLE estimator is asymptotically efficient. Furthermore, MLE estimator is always sufficient.

4.3 The Bayesian Approach

Prior distribution can be used as "add-one smoothing". The posterior distribution can be treated as weighted average of two estimators: 1) estimator without considering priors, and 2) estimator when there are no data.

The MSE Ratio of Bayesian estimator and MLE estimator approaches 1 as sample size approaches infinity. Whether it is larger or smaller than 1 depends on the prior and true parameters.

Conjugate prior can be used for computing efficiency issues. For now since the increase of computation powers, it is of less interested.

5 Testing Hypotheses and Assessing Goodness of Fit

5.1 Terminology

Type I error: Rejecting H_0 when it is true; $\alpha = P(TypeIerror) = P(RejectH_0|H_0)$.
Type II error: Accepting H_0 when it is false; $\beta = P(TypeIIerror) = P(RejectH_0|H_A)$. Power of the test $= 1 - \beta$.
Simple Hypotheses: hypotheses that completely specifies the probability distribution; Composite Hypotheses: hypotheses that do not completely specifies the probability distribution.

P-value is defined as the probability from the observed point to the right under the null distribution. It is actually *the smallest value of α for which the null hypothesis will be rejected*, which means, if the the p-value is larger than α the test will not reject, otherwise it will reject.

The goal of the test is to achieve low α and high $1 - \beta$. Obviously it is hard to achieve these two goals at the same time. One way to resolve this trade-off is to fix α in advance, and try to maximize $1 - \beta$.

How to decide which is null and which is alternative hypotheses? 1) Choose the simpler of two hypotheses as the null; 2) Choose the one when falsely rejecting it may cause more serious consequences.

5.2 The Neyman Pearson Lemma

Neyman-Pearson Lemma told us that, among all possible tests achieving significance level $\leq \alpha$ (which means, both the hypotheses are simple), the test based on *likelihood ratio* $\frac{f_0(x)}{f_A(x)}$ maximizes the power.

However, in real situations we are seldom presented with the problem of testing two simple hypotheses. Therefore we use the concept of *uniformly most powerful* for composite tests, it exists for some common one-sided alternatives.

5.3 Duality of Confidence Intervals and Hypothesis Tests

The hypothesis test's parameter lies in the confidence interval if and only if the hypothesis test accepts. This duality can be quite useful when one of them is hard to derive but the other one is easy.

5.4 Generalized Likelihood Ratio Tests

The generalized likelihood ratio $\Lambda = \frac{\max_{\theta \in \omega_0}[lik(\theta)]}{\max_{\theta \in \Omega}[lik(\theta)]}$.

Theorem 5.1 *Under smoothness conditions on the probability density or frequency functions involved, the null distribution of -2logΛ tends to a chi-*

square distribution with degrees of freedom equal to dimΩ - dimω_0 as the sample size tends to infinity.

For Multinomial Distribution where θ is a vector of k parameters to be estimated, we need to know whether the model $p(\theta)$ is good or not, according to the observed data (cell counts). We can derive $\Lambda = \frac{\max_{\theta\in\omega_0}[lik(\theta)]}{\max_{\theta\in\Omega}[lik(\theta)]}$

$= \Pi(\frac{p_i(\hat{\theta})}{\hat{p_i}})^{x_i}$

$= \Pi(\frac{p_i(\hat{\theta})}{\hat{p_i}})^{np_i}$

Then $-2\log\Lambda = -2n\Sigma\hat{p_i}\log(\frac{p_i(\hat{\theta})}{\hat{p_i}})$

$= 2\Sigma n\hat{p_i}\log(\frac{n\hat{p}_i}{np_i(\hat{\theta})})$

$= 2\Sigma O_i\log(\frac{O_i}{E_i})$

Where O_i is the observed counts and E_i is the expected counts. The $-2\log\Lambda$ is asymptotically χ^2_s with degrees of freedom s = dimΩ - dimω_0 = $(m-1)-k$. Where k = length of the vector θ = number of parameters in the model. This generalized likelihood ratio test is called G^2.

The Pearson's Chi-square test $X^2 = \Sigma\frac{[O_i-E_i]^2}{E_i}$. G^2 and X^2 are asymptotically equivalent Under H_0 under H_0. This can be proved by Taylor expansions. It appears G^2 test should be "more accurate", but X^2 is actually more frequently used since it is somewhat easier to calculate without the use of a computer.

If one has a specific alternative hypothesis in mind (which mean, the cell probabilities are not completely free), better power can usually be obtained.

6 Summarizing Data

6.1 Empirical cumulative distribution function

We can show that the empirical cumulative distribution function has a binomial distribution, which can give us the estimated variance of the function.

6.2 The Survival Function

The survival function is equivalent to the CDF: $S(t) = 1 - F(t)$.

6.3 Quantile-Quantile (Q-Q) Plots

Given the samples $x_1, ..., x_n$, they can be viewed as the $\frac{k}{n+1}th$ empirical quantile. We add one to the denominator since we can never get the 100%

quantile. Q-Q Plots can be used to test the fitness of the model, by plotting samples with order statistics $x_{(k)}$ and theoretical quantiles $y_{(\frac{k}{n+1})}$. If F_X is the true distribution, the Q-Q plot will likely be close to a 45^O straight line. Q-Q Plots can also be used for analyzing relations between variables, especially when the samples of the two variables are observed separately.

6.4 Measure of Location

From many measurements of locations, arithmetic mean is sensitive to the large values of samples, called "outliers". The median is much more "robust" than the mean, but not efficient as the mean if the data are free of outliers (sorting is more time consuming than summing up.) Trimmed Mean $\frac{\{}{x}(n\alpha + 1) + ... + x_{\{}(n - [n\alpha])\}\{n - 2[n\alpha]\}\}$ is more robust than the mean and less sensitive to outliers.

7 Comparing Two Samples

There are two cases where we would like to compare two samples: 1) compare the independent samples, where we usually has some hypothesis with their parameter correlations; 2) compare the paired samples,

Comparing two independent samples are usually used when we want to argue that two "methods" (e.g. treatment to patients, randomized algorithm to problems, etc) have same effects (same distribution parameters), or one is better than the other (one-sided test). And we usually use the *t-test*:

$t = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \backsim t_{m+n-2}$

Where $s_p^2 = \frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}$. There is a equivalence between t-test and Likelihood Ratio Test.

Under different alternative hypothesis choices, we have different power analysis. In general, three parameters, confidence α (larger value results in larger power), distance between the null and alternative hypothesis $\triangle$ (larger value results in larger power), and sample size n, affect the power (larger value results in larger power). Note the power analysis is based on the assumption that alternative hypothesis illustrate the real distribution ($\triangle$, $\frac{2\sigma^2}{n}$), instead of N(0, $\frac{2\sigma^2}{n}$). But we will reject when $\bar{X}-\bar{Y} > z_{\alpha/2}\sigma\sqrt{\frac{2}{n}}$ (under the null hypothesis). When computing the power we need to transform this expression under the assumption of alternative hypothesis.

Comparing paired samples requires considering the correlation between two variables. If there is positive correlation the variance of the estimator is smaller, and vice versa.

8 Linear Least Squares

8.1 The Basic Procedure

Given the *observed* data points (x_i, y_i), and assume linear relationship $y = \beta_0 + \beta_1 x$ between y and x. Estimate the linear parameters β_0, β_1 by minimizing the mean square error (MSE) $\sum(y_i - \beta_0 - \beta_1 x_i)^2$:
$\hat{\beta}_0 = E(Y) - E(X)\hat{\beta}_1, \hat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)}$.

8.2 Conditional Expectation and Prediction

If X and Y are jointly normal *variables*, then linear regression is the best estimator under MSE criterion.

8.3 Best Linear Estimator

Observing X, to predict Y. Assume Y = a + bX. Find the optimal a and b for achieving the smallest MSE. The parameter estimators are the same as the basic procedure, besides the distribution parameters μ, σ, ρ might be known.