36-705 – Intermediate Statistics.

Key areas to understand.

A pedagogical tool to track the key aspects of the lecture that the instructor emphasises, and as a checklist of things you have judged are important.

List what you feel a need to commit to memory later.

Week 1

Lecture Notes 1 – Review of probability.

YouTube lecture 31/08/2016.

- There was a lecture that covered earlier parts of the probability review that is not covered here.
- Understand that independence can be a physical fact, or an assumption that has to be evaluated.
- Distinction between parameter and random variable.
- The importance of the IID data to the statistical setting.
- Distinction between the distribution of IID random variables, and the (sampling) distributions of statistics/estimators of IID random variables.
- Source of stochasticity in sampling estimators, and their characterisation as random variables/deterministic functions of random variables.
- Properties of the mean and variance of sample mean and variance estimators i.e. properties of the sampling distribution, and their relation to the mean and variance of the underling IID random variables.

Lecture Notes 2 – Inequalities.

YouTube lecture 31/08/2016.

- The proof of Lemma 4 (as part of a strategy to prove Hoeffding) is not in the video footage.
- Appreciate role of probability bounds in machine learning, particularly Hoeffding's inequality.
- Probability bounds -> VC theory -> convergence bounds.
- Understand the proofs of the Gaussian tail, Markov, Chebyshev inequalities.

YouTube lecture 02/09/2016.

- Understand the proofs of Hoeffding's lemma and Chernoff's method.
- Understand the proof of Hoeffding's inequality, the variational trick.
- Understand role of Cauchy-Schwarz and Jensen's inequality for to bound expectations.
- Application to Kullback-Leibler divergence as having certain desirable distance metric properties.
- Understand the proof of the bound on the expectation of the maximum of a series of IID random variables.
- Appreciate the relation between a "thin-tail" in a probability distribution and an exponential bound on the corresponding moment-generating function – as a sub-Gaussian random variable.

Week 2.

YouTube lecture 07/09/2016.

- Understand the computer-science definitions of little-o and Big-O notation.
- Understand the adaptation of these definitions into a statistical/probabilistic context i.e. little-op and Big-Op.
- Understand that they are both notions of convergence in probability and stochastic boundedness.
- Understand and be able to use arguments from mathematical analysis and probability inequalities to prove composite statements about random variables stipulated in this notation.

Lecture Notes 3 – Uniform bounds.

- Appreciate the limitations of Markov's, Chebyshev's and Hoeffding's inequality as bounds on one random variable.
- Appreciate that many applications in statistics and machine learning require bounds on multiple random variables.
- Understand the definition of the empirical CDF and theoretical CDF.
- Understand the distinction between pointwise and uniform convergence.
- Understand Hoeffding's inequality as a statement about pointwise convergence.
- Understand that uniform convergence can be viewed as a probability bound on the maximum approximation error.
- Understand that uniform convergence is a stronger statement that pointwise convergence.
- Appreciate the role of uniform convergence in computer science, statistical, and machine learning settings.
- Understand this distinction in context of estimation error on CDF and of training error in classification.

Week 3.

YouTube lecture 12/09/2016.

- Understand the statement of uniform bounds over finite classes of sets, or of functions, and its derivation.
- Understand the role of the complexity of the class of sets.
- Understand shattering, and the shatter coefficient.
- Understand how to apply these concepts to familiar classes of sets.
- Understand the theorem due to Vapnik-Chervonenkis giving a uniform bound over a class of (possibly infinite) sets in terms of an exponential and shattering coefficient.
- Understand the relation between the shattering coefficient and VC dimension.
- Understand Sauer's theorem on the behaviour of shattering coefficient and its relation with the VC dimension.

Lecture Notes 4 – Convergence theory.

- Understand the notion of a statistic.
- Understand the distinction between properties of a sequence of random variables, and properties of a sequence of statistics.
- Appreciate the distinction between convergence in the setting of mathematical analysis, and in the probabilistic setting.
- Understand the following probabilistic formulations of convergence: almost sure convergence, convergence in probability, convergence in quadratic mean, and convergence in distribution.

- Understand that the main preoccupations of the course are convergence in probability, and in distribution.
- Understand the taxonomy of relationships between these formulations of convergence.
- Understand how to prove convergence in distribution.

YouTube lecture 14/09/2016.

- Understand how to prove convergence in probability.
- Understand the directions of implication in the taxonomy of convergences.
- Understand the proof for convergence in quadratic mean implying convergence in probability.
- Understand the proof for convergence in probability implying convergence in distribution.
- Understand the proof that convergence in probability does not imply convergence in quadratic mean, by counterexample.
- Understand that convergence in quadratic mean is a relatively stronger statement about the moments and tails of a distribution.
- Understand the proof that convergence in distribution does not imply convergence in probability.
- Understand that convergence in probability is a statement about random variables, whilst convergence in distribution is about the probability statements one makes about random variables.
- Understand how convergence in distribution can be combined with rescaling to give meaningful approximations in situations where working with the existing probability distribution is intractable.
- Understand the theorems concerning preservation of convergence under transformations, Slutsky's theorem, and the continuous mapping theorem.
- Understand that the LLN and CLT are statements about convergence in probability and in distribution respectively.
- Understand the statement and proof of the Weak Law of Large Numbers (WLLN) for finite variance.
- Understand the statement of the Strong Law of Large Numbers (SLLN).

YouTube lecture 16/09/2016.

- Understand the significance and applicability of the Central Limit Theorem (CLT).
- Understand that generalisations of the CLT exist for non-IID settings.
- Understand a simplified proof of the CLT using moment-generating functions.
- Understand the lemma that the convergence of a sequence of moment generating functions can be used to show convergence in distribution.
- Understand the Berry-Esseen theorem.
- Understand that the Berry-Esseen theorem specifies the rate of convergence of a probability of a scaled mean of a random sample to a Normal distribution asymptotically.
- Understand that the Berry-Esseen theorem provides a bound on the maximal error of approximation between the Normal distribution and the true distribution of the scaled sample mean.
- Understand the proof of the CLT in which the sample variance is used.
- Understand how theorems concerning preservation of convergence under transformations (e.g. continuous mapping, Slutsky, etc.) are invoked as proof strategies.

YouTube lecture 19/09/2016.

- Understand the extension of the CLT to the multivariate case.
- Understand that IID random vectors are vectors containing random variables.
- Understand that independence for IID random vectors is across random vectors, not within the elements of a single random vector.
- Understand the motivation for the use of the Delta method, as a means of finding the limiting distribution of a smooth function of a variable that has a limiting Normal distribution.
- Understand the heuristic derivation of the Delta method via Taylor series approximations.
- Understand the extension of the Delta method to the multivariate case.

Lecture Notes 5 – Statistical inference.

- Understand the differences in mode of enquiry in a probabilistic setting, in contrast to a statistical setting.
- Understand the definition of a statistical model as a set of possible probability distributions that could have generated the data.
- Understand that a distinction is made between parametric and non-parametric models, via whether a model can be specified by a finite number of parameters.
- Understand the significance of sufficiency for statistics in the pre-computer era, computer era, and Big Data era.
- Understand sufficiency through data reduction and information loss.
- Understand the formal definition of a sufficient statistic.
- Understand that in terms of data reduction, or information loss, "sufficiency" may not be a sufficient criterion for a "good statistic".

YouTube lecture 21/09/2016.

- Understand how the definition of a sufficiency in terms of parameter dependence can be used to check if a statistic is sufficient.
- Understand sufficiency through partitions of the sample space, and counterexamples.
- Understand the results governing the equivalence of statistics inducing the same partition, and sub-partitions.
- Understand the factorisation theorem.
- Understand how the factorisation theorem may be used to evaluate whether a statistic is sufficient.
- Understand the formal definition of a minimal sufficient statistic.
- Understand that minimal sufficient statistics generate the coarsest sufficient partition.
- Understand that minimal sufficient statistics can be expressed as functions of other sufficient statistics.
- Understand that from the perspective of partitioning the sample space, a sufficient statistic can be viewed as inducing further sub-partitions of the partition induced by a minimal sufficient statistic.
- Understand the theorem that allows one to assess whether a sufficient statistic is also minimally sufficient.
- Understand that the minimal sufficient statistic is not unique, but the minimal sufficient partition is unique.
- Understand the nuance that the sufficiency of a statistic means that you have all the information needed to compute the likelihood function, rather than the sufficient statistic containing all the information in the data.

Week 5.

YouTube lecture 26/09/2016.

Lecture Notes 6 – The likelihood function.

- Understand the role of the likelihood function in generating (maximum likelihood) estimators, and in Bayesian inference.
- Understand the formal definition of the likelihood function.
- Understand the distinction between a probability density function and the likelihood function in terms of parameters and data.
- Understand that likelihood functions are equivalent if they are proportional.
- Understand that if likelihood functions of two data-sets are proportional, then those two data-sets lie in the same partition of a sample space.
- Understand the relation between sufficient statistics, minimal sufficient statistics, and the likelihood function through partitions of the sample space.

Lecture Notes 7- Point estimation.

- Understand the problem setup for point estimation of parametric models – estimation of parameters given data.
- Understand the formal definition of an estimator as a function of data, and also as a random variable.
- Reinforce the distinction between a parameter and an estimator.
- Understand the method of moments (MoM), maximum likelihood estimation (MLE), and Bayesian estimators, as procedures for generating estimators.
- Be aware of brief contexts of each, as well as recent trends regarding the resurgence in the use of MoM procedures.
- Understand the distinction between:
    - Procedures for generating estimators.
    - Criteria for evaluating the quality of estimators.
- Understand the concerns of each criterion for evaluating estimators (consistency, bias-variance, mean-squared-error, minimax theory, robustness).
- Understand the terminology, formal definitions of the expectation of an estimator, bias, sampling distribution of an estimator, standard error, consistency.

YouTube lecture 28/09/2016.

- Understand how the MoM procedure generates estimators as solutions to the equations of sample moments with theoretical moments.
- Understand how the ML procedure generates estimators as solutions to the maximisation of the likelihood function.
- Understand that maximisation occurs over the log-likelihood function rather than the log-likelihood function, for reasons of tractability.
- Understand that in real-world settings, numerical methods will be required to compute maximum likelihood estimates.
- Understand the formal definition of the profile likelihood function.
- Understand the geometric interpretation of the profile likelihood via iso-contours.
- Understand that the maximiser of the likelihood and profile likelihood are equivalent.
- Understand the property of equivariance of MLE.
- Understand that equivariance can be invoked to easily derive estimators of transformations of parameters, and that these transformations need not be monotonic.

- Understand the instructor's emphasis that Bayesian estimators, in particular, are an algorithm for generating estimators.
- Understand that the above procedure involves the treatment of the parameter as a random variable, the construction of prior distribution, and the computation of a posterior, using Bayes theorem.
- Understand the Bayesian estimator generates estimators by taking the mean of a posterior distribution over the parameter.
- Understand the notion of a conjugate prior.
- Understand the mode of a posterior distribution over parameters gives the maximum a posterior (MAP) estimator.

YouTube lecture 30/09/2016.

- Understand the formal definition of mean-squared error (MSE).
- Understand the theorem and proof relating MSE to bias and variance.
- Understand that MSE provides a metric for evaluating estimators, but does not give clear prescription on how this can be used to select estimators from a subset of estimators.
- Understand that 'best unbiased estimators' refer to an estimator with the smallest variance out of a class of unbiased estimators.
- Understand the Rao-Blackwell theorem.
- Understand that Rao-Blackwell theorem provides a way of unilaterally decreasing the variance of an unbiased estimator, by computing the mean, conditioned on a sufficient statistic.

Lecture Notes 8 – Minimax Theory.

- Understand that minimax theory provides a theoretical framework to evaluate the 'quality' of an estimator.
- Understand that minimax formalises the quality of an estimator through a loss function.
- Understand that the loss function encodes an idea of distance between an estimator from the true parameter value.
- Be aware of the variety of functional forms the loss function can take.
- Understand the formal definition of the risk of an estimator as the expected value of the loss.
- Understand that in context of minima theory, MSE is the risk of an estimator under squared-error or L2 loss.
- Understand the formal definition of minimax risk.
- Understand that comparison of risk functions of estimators often does not provide conclusive results on which is a more desirable estimator.
- Understand that the maximum risk and Bayes risk can provide a systematic way of comparing a set of estimators.
- Understand the formal definition of the Bayes risk.
- Understand that the Bayes estimator minimises the Bayes risk.
- Understand that the minimax estimator minimises the maximum risk; or achieves minimax risk.

Week 6.

YouTube lecture 03/10/16.

Lecture unexpectedly terminates after 20 minutes. The material that was assumed to be covered in the remaining 30 minutes, is listed in italics.

- Understand that minimisation of the Bayes risk is often used as a means of computing minimax estimators.
- Understand the formal definition of the posterior risk.
- Understand that the posterior risk is an expectation of the loss function with respect to a posterior density (as opposed to with respect to a sampling distribution).
- Understand the theorem of and proof for the result that the Bayes estimator minimises the posterior risk, and as a consequence, the Bayes risk.
- Understand that this equivalence provides a more tractable means of computing the Bayes estimator.
- Understand that the Bayes estimator is the:
    - Mean of the posterior distribution (over the parameter) under L2 loss.
    - Median of the posterior distribution under absolute error loss.
    - Mode of the posterior distribution under zero-one loss.
- *Insert material here after your own review.*

<u>YouTube lecture 05/10/2016.</u>

- Understand the proof of the sample means as the minimax estimator under L2 loss, Normally distributed IID random vectors.
- Appreciate that the proof strategy consists of artfully selecting the Bayes risk and maximum risk to upper and lower bound the minimax risk.
- Understand the proof as an illustrative example of a central preoccupation in statistical decision theory, that is:
    - Finding the minimax risk.
    - Finding an estimator that achieves minimax risk i.e. the minimax estimator.
- Understand that under certain regularity conditions, the MLE is approximately minimax.

<u>Lecture Notes 9 – Asymptotic Theory.</u>

- Reinforce understanding of convergence, boundedness, convergence in probability, and stochastic boundedness.
- Be aware of the various distance metrics between probability distributions e.g. total variation, L1-distance, Hellinger distance, KL-divergence.
- Understand some of the properties of these distance metrics.
- Understand the formal definitions of consistency and asymptotic Normality.
- Understand that both definitions are concerned with the asymptotic behaviour of an estimator i.e. as more samples are drawn.
- Understand that an estimator can be shown to be consistent directly by proof that the estimator converges in probability to the true parameter, or indirectly, via proof of convergence in quadratic mean.
- Understand that this is equivalent to proving that the MSE converges to 0.
- Understand the role of regularity conditions in the consistency of the MLE.
- Understand the theorem and proof for the result that the probability of the likelihood at the true value of the parameter being greater that of the likelihood for an arbitrary parameter approaches 1 asymptotically.
- Understand the role of the identifiability regularity condition in the above.

<u>YouTube lecture 07/10/2016.</u>

This lecture (shown as Lecture 14 – Asymptotics, 7th October 2016 on the course page) is missing on the YouTube playlist.

The contents of what was covered in this lecture needs to be inferred.

Insert material here after review.

Week 7.

YouTube lecture 10/10/16.

- Understand that in order to show consistency of the MLE, additional work beyond the previous result that is beyond the scope of the course is required.
- Understand that this additional work lies in the distinction between pointwise and uniform convergence.
- Understand the result that MLE is consistent under the following regularity conditions: fixed parameter dimensionality, smoothness of the probability density function, parameter identifiability.
- Understand the formal definitions of the score function and Fisher information.
- Have how the above is related graphically.
- Understand that the variance of the MLE is the inverse of the Fisher information.
- Understand the interpretation of statistical "information" in this context.
- Understand the theorem and proof that the expected value of the score function is 0, under appropriate regularity conditions.
- Understand that consequently, the Fisher information is the 2^{nd} moment and variance of the score function.
- Understand the relation between the Fisher information for IID data of an arbitrary sample size, and for a one-data point sample size.
- Understand that the theorem and proof that the Fisher information is the negative expected value of 2^{nd} derivative of the likelihood, under regularity conditions.
- Understand generalisations of the score function and the Fisher information to the score vector function and the Fisher information matrix.
- Understand the formal definition of asymptotic Normality.
- Understand the proof of asymptotic Normality for the MLE (condensed).

YouTube lecture 12/10/16.

- Review lecture – tbc.

Week 8.

YouTube lecture 17/10/16.

- Understand the proof of asymptotic Normality for the MLE (detailed).
- Understand how Fisher information and asymptotic Normality can be used to approximate standard errors.
- Understand how in practice, standard errors are estimated using the Fisher information evaluated at MLE.
- Understand that under certain conditions on the Fisher information, the continuous mapping theorem renders estimators of the standard error consistent.
- Understand how the delta method can be used to construct standard errors of functions of parameters.
- Understand how the delta method can be used to construct estimators of standard errors of functions of MLE.
- Understand that "well-behaved" estimators (finite dimensional models under regularity conditions) are asymptotically Normal.

- Understand that under technical conditions, MLE is "optimal" in the sense of statistically efficient.
- Understand that statistically efficient estimators have the smallest possible asymptotic variance.
- Understand the formal definition of asymptotic relative efficiency (ARE) as the ratio of variances.
- Understand that ARE provides a way of comparing well-behaved/asymptotically Normal estimators.
- Understand that efficiency of MLE means that ARE of the MLE with an arbitrary estimator is less than 1.
- Understand the multivariate vector generalisation of asymptotic Normality, approximate standard errors, and delta method.
- Understand the formal specification of the exponential family of distributions.
- Understand that efficiency of MLE is contingent on model correctness (in a parametric model framework).
- Understand that literature of robust statistics concerns trade-offs between robustness and efficiency.
- Understand that the motivation for non-parametric methods can be viewed as a means of addressing model correctness issues.

YouTube lecture 19/10/16.

Lecture Notes 10 – Hypothesis Testing.

- Understand the nature of settings in which hypothesis testing under a parametric framework is appropriate.
- Understand the formal setup of a hypothesis test.
- Understand the distinction between null and alternate hypotheses under the general setting of parameter spaces.
- Understand the distinction between simple and composite null hypotheses.
- Understand the distinction between type I (false positive) and type II errors (false negative).
- Understand that hypothesis testing is often conducted in an "asymmetric situation".
- Understand that this asymmetric situation consists of type I errors being less tolerable than type II errors.
- Understand that the general strategy for hypothesis test design consists of controlling for an acceptable false positive rate as a constraint, and then minimising the false negative rate.
- Understand that acceptable false positive rates vary according to context and the epistemological requirements of a scientific discipline.
- Understand that constructing hypothesis tests consists of appropriate selection of a test statistic, rejection region.
- Understand that desirable statistical properties of hypothesis tests motivate appropriate choices of test statistic and rejection region.
- Understand the formal definition of a power function.
- Understand that the general strategy for test design in terms of type I error rates, and constrained optimisation of the power function over distinct domains.
- Understand the distinction between size and level.
- Understand the distinction between one-sided and two-sided alternatives.
- Understand the interpretation of the slope and convergence of the power function statistically.
- Understand the definition of a critical value.

- Understand the formal specification of the Neyman-Pearson test.
- Understand that the Neyman-Pearson is the uniformly most powerful (UMP) level-α test.
- Understand that the Neyman-Pearson test is in practice, limited in applicability.