Lecture 5

Classical Statistical Inference

* Three main concepts in statistical inference:

1) Point estimate: what is the best estimate for a model parameter Θ.
2) Confidence estimate: How confident should we be in that estimate?
3) Hypothesis testing: Are the data consistent with the model used?

Statistical Paradigms

Classical (frequentist) paradigm

- Parameters, Θ, are fixed, unknown constants
- Probabilities refer to relative frequencies of events.
- Statistical procedures should have well-defined long-run frequency property.

Bayesian paradigm

- Data is fixed, known constant.
- Probability describes subjective belief in model & parameters, Θ.
- Inference about a parameter is by producing a pdf, $p(\Theta)$.

eg: a 95% confidence interval should bracket the true value with a limiting frequency of at least 95%.
- Can be computationally cheap/easy.

This pdf quantifies the uncertainty about the parameter & can be used to make point estimate
- Usually more computationally demanding.

- The important thing to note that the best-fit solutions are typically consistent between frequentist & Bayesian methods.

Maximum Likelihood Estimation

- MLE is a common special case of frequentist analyses.
- However, Bayesian analyses build on top of the MLE framework too.

- Note: frequentist analyses, unlike Bayesian analyses, are not tied to the likelihood.

- You can use other cost functions in frequentist analyses, eg. MSE, ELU, RELU.

- The data "likelihood" represents a quantitative description of our measuring process.
eg. if we know (or assume) that our data $\{x_i\}$ are drawn from $N(\mu, \sigma)$, then the likelihood for any single data point is

$$p(x_i \mid \mu(\vec{\theta})) = N(\mu, \sigma)$$

$p(x_i | M(\theta)) = N(\mu, \sigma)$

- For the full dataset, the likelihood is given by a product of individual probabilities /lk/:

$$\mathcal{L} \equiv p(\{x_i\} | M(\vec{\theta})) = \prod_{i=1}^{N} p(x_i | M(\vec{\theta}))$$

- If our data is drawn from a Gaussian, then $M(\vec{\theta}) \equiv [\mu, \sigma]$.

- Note that the likelihood, $\mathcal{L}$, is not a true (properly normalized) pdf, nor is it a statement on the probability of obtaining model parameters, $\vec{\theta}$.

- Since $\mathcal{L}$ is a product of probabilities, all < 1, the $\mathcal{L}$ can be a very small number, and we instead work with $\ln \mathcal{L}$ instead.

- The maximum likelihood approach can be broken down into the following steps:

1. Formulate the data likelihood, $p(D|M)$, for some model $M(\vec{\theta})$ with parameters $\vec{\theta}$.
2. Search for the parameters, $\vec{\theta}^{\circ}$, that maximize $p(D|M)$.
   - These parameters, $\vec{\theta}^{\circ}$, are the point estimates.
3. Determine the confidence region for model parameters $\vec{\theta}^{\circ}$:
   - either analytically with the Fisher matrix if $\mathcal{L}$ is well-behaved.

- numerically by either bootstrap, jack-knife or cross-validation.

④ Perform hypothesis testing to see if your chosen model explains the observed data. If not, find a new $M'(\vec{\theta})$ and go back to ①.

- Properties of ML estimators:
  - They are consistent estimators, i.e. they can be proven to converge to the true parameter value as number of data points increases.
  - They are asymptotically normal estimators, i.e. as number of data points increase, the parameter distribution approaches a normal distribution centered on the MLE.
  - They asymptotically achieve the theoretical minimum possible variance, called the Cramer-Rao bound.

- Thus, to summarize, given a likelihood, $\mathcal{L}$, the MLE, $\vec{\theta}^\circ$, is given by:

$$\boxed{\left.\frac{d \ln \mathcal{L}(\vec{\theta})}{d\vec{\theta}}\right|_{\vec{\theta}^\circ} = 0}$$

- Using the asymptotic normality of MLE, the uncertainty matrix for these parameters can be derived by,

$$\boxed{\sigma_{jk} = \left([F^{-1}]_{jk} \right)^{1/2}}$$

where,

$$\boxed{F_{jk} = -\frac{d^2 \ln \mathcal{L}}{d\theta_j \, d\theta_k}\bigg|_{\theta = \theta_0}}$$