

Lecture 2. Statistical Schools of Thought

COMP90051 Statistical Machine Learning

Semester 2, 2020
Lecturer: Ben Rubinstein



THE UNIVERSITY OF
MELBOURNE

This lecture

How do learning algorithms come about?

- Frequentist statistics
- Statistical decision theory
- Extremum estimators
- Bayesian statistics

Types of probabilistic models

- Parametric vs. Non-parametric
- Generative vs. Discriminative

Frequentist Statistics

Wherein unknown model parameters are treated as having fixed but unknown values.

Frequentist statistics

- Abstract problem

- * Given: X_1, X_2, \dots, X_n drawn i.i.d. from some distribution
- * Want to: identify unknown distribution, or a property of it

Independent and
identically distributed

- Parametric approach (“**parameter estimation**”)

- * Class of **models** $\{p_\theta(x): \theta \in \Theta\}$ indexed by **parameters** Θ (could be a real number, or vector, or)
- * **Point estimate** $\hat{\theta}(X_1, \dots, X_n)$ a function (or **statistic**) of data

Hat means estimate
or estimator

- Examples

- * Given n coin flips, determine probability of landing heads
- * Learning a classifier

Estimator Bias

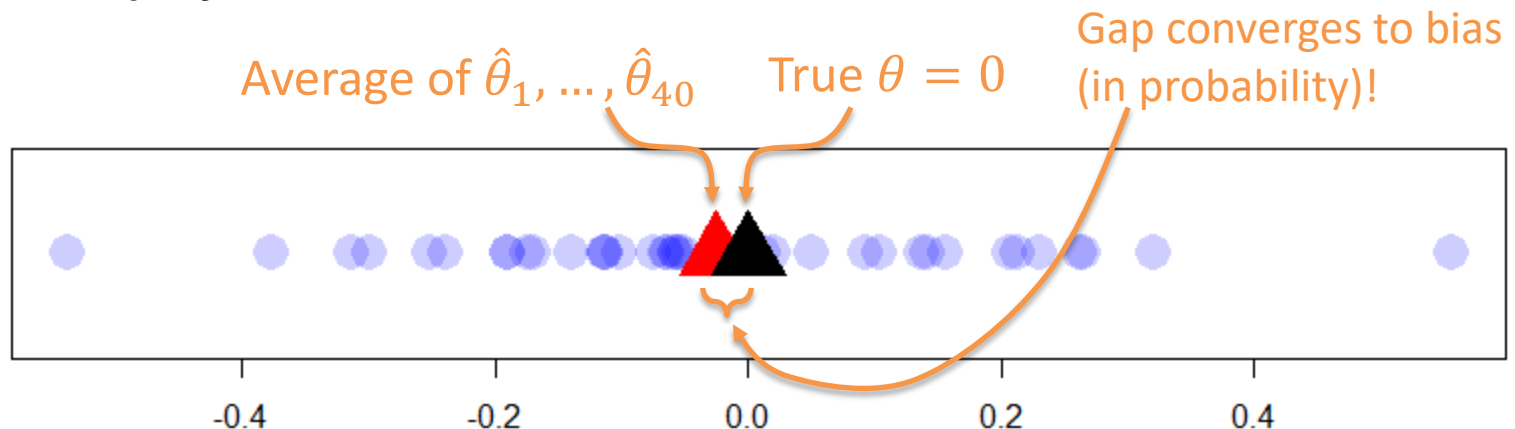
Frequentists seek good behaviour, in ideal conditions

- **Bias:** $B_{\theta}(\hat{\theta}) = E_{\theta}[\hat{\theta}(X_1, \dots, X_n)] - \theta$

Subscript θ means data really comes from p_{θ}

Example: for $i=1\dots 40$

- $X_{i,1}, \dots, X_{i,20} \sim p_{\theta} = \text{Normal}(\theta = 0, \sigma^2 = 1)$
- $\hat{\theta}_i = \frac{1}{20} \sum_{j=1}^{20} X_{i,j}$ the sample mean, plot as ●



Estimator Variance

Frequentists seek good behaviour, in ideal conditions

- **Variance:** $\text{Var}_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2]$

$\hat{\theta}$ still function of data

Example cont.

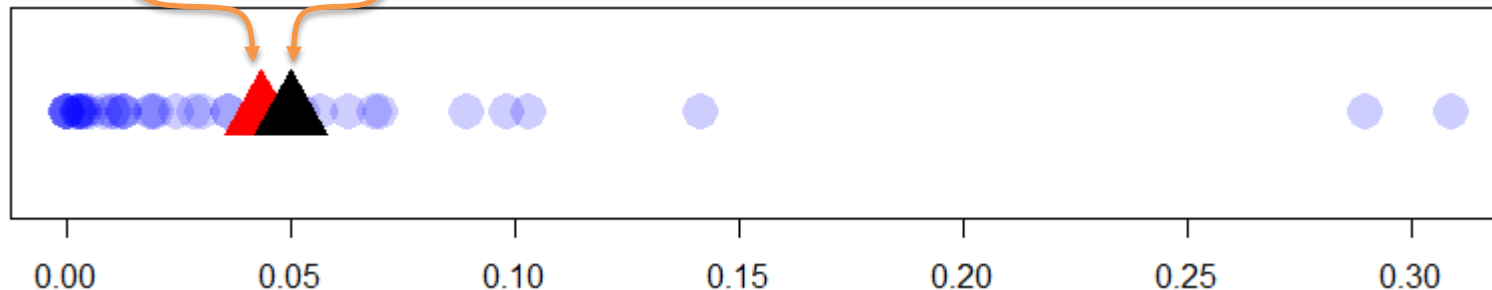
- Plot each $(\hat{\theta}_i - \mathbb{E}_\theta[\hat{\theta}_i])^2 = \hat{\theta}_i^2$ as



Average of $\hat{\theta}_1^2, \dots, \hat{\theta}_{40}^2$

True $\text{Var}_\theta(\hat{\theta}) = \frac{\sigma^2}{20} = 0.05$

Once again, average converges to true (in probability)!



Asymptotically Well Behaved

For our example estimator (sample mean), we could calculate its exact bias (zero) and variance (σ^2). Usually can't guarantee low bias/variance exactly 😞

Asymptotic properties often hold! 😊

Bias closer and closer to zero

- **Consistency**: $\hat{\theta}(X_1, \dots, X_n) \rightarrow \theta$ in probability
- **Asymptotic efficiency**: $\text{Var}_{\theta}(\hat{\theta}(X_1, \dots, X_n))$ converges to the smallest possible variance of any estimator of θ

Variance closer & closer to optimal

Amazing Cramér-Rao lower bound (**outside subject scope**):

$\text{Var}_{\theta}(\hat{\theta}) \geq \frac{1}{I(\theta)}$ with $I(\theta)$ the Fisher information of p_{θ} for any $\hat{\theta}$

Maximum-Likelihood Estimation

- A **general principle** for designing estimators
- Involves **optimisation**
- $\hat{\theta}(x_1, \dots, x_n) \in \operatorname{argmax}_{\theta \in \Theta} \prod_{i=1}^n p_{\theta}(x_i)$
- *“The best estimate is one under which observed data is most likely”*



Fischer

Later: MLE estimators usually well-behaved asymptotically

Example I: Bernoulli

- Know data comes from Bernoulli distribution with unknown parameter (e.g., biased coin); find mean

- MLE for mean $\mathcal{L}(\theta) = \log \prod_{i=1}^n p_{\theta}(x_i) = \bar{X} \log \theta + (n - \bar{X}) \log(1 - \theta)$
 $\bar{X} = \sum_{i=1}^n X_i$

$$* p_{\theta}(x) = \begin{cases} \theta, & \text{if } x = 1 \\ 1 - \theta, & \text{if } x = 0 \end{cases} = \theta^x (1 - \theta)^{1-x}$$

(note: $p_{\theta}(x) = 0$ for all other x)

- Maximising likelihood yields $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i$

$$\frac{d}{d\theta} \mathcal{L}(\hat{\theta}) = \frac{\bar{X}}{\hat{\theta}} - \frac{n - \bar{X}}{1 - \hat{\theta}} = 0 \Rightarrow \hat{\theta} = \frac{\bar{X}}{n}$$

Example II: Normal

- Know data comes from Normal distribution with variance 1 but unknown mean; find mean

- MLE for mean

- * $p_{\theta}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \theta)^2\right)$

- * Maximising likelihood yields $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i$

- Exercise: derive MLE for *variance* σ^2 based on

$$p_{\theta}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \text{ with } \theta = (\mu, \sigma^2)$$

MLE 'algorithm'

1. Given data X_1, \dots, X_n **define** probability distribution, p_θ , assumed to have **generated the data**
2. Express likelihood of data, $\prod_{i=1}^n p_\theta(X_i)$
(usually its **logarithm... why?**)
3. Optimise to find *best* (most likely) parameters $\hat{\theta}$
 1. take partial derivatives of log likelihood wrt θ
 2. set to 0 and solve
(failing that, use **gradient descent**)

Mini Summary

- Frequentist school of thought
- Point estimates
- Quality: bias, variance, consistency, asymptotic efficiency
- Maximum-likelihood estimation (MLE)

Next: Statistical Decision Theory, Extremum estimators

Statistical Decision Theory

Branch within statistics, optimisation, economics, control, emphasising utility maximisation.

Decision theory



Wald

- Act to maximise utility - connected to economics and operations research
- **Decision rule** $\delta(\mathbf{x}) \in A$ an action space
 - * E.g. Point estimate $\hat{\theta}(x_1, \dots, x_n)$
 - * E.g. Out-of-sample prediction $\hat{Y}_{n+1} | X_1, Y_1, \dots, X_n, Y_n, X_{n+1}$
- **Loss function** $l(a, \theta)$: economic cost, error metric
 - * E.g. square loss of estimate $(\hat{\theta} - \theta)^2$
 - * E.g. 0-1 loss of classifier predictions $1[y \neq \hat{y}]$

Risk & Empirical Risk Minimisation (ERM)

- In decision theory, really care about *expected* loss
- **Risk** $R_{\theta}[\delta] = \mathbb{E}_{\mathbf{X} \sim \theta}[l(\delta(\mathbf{X}), \theta)]$
 - * E.g. true test error
 - * aka generalization error
- Want: Choose δ to minimise $R_{\theta}[\delta]$
- Can't directly! Why?
- **ERM**: Use training set \mathbf{X} to approximate p_{θ}
 - * Minimise **empirical risk** $\hat{R}_{\theta}[\delta] = \frac{1}{n} \sum_{i=1}^n l(\delta(X_i), \theta)$

Decision theory vs. Bias-variance

We've already seen

- Bias: $B_{\theta}(\hat{\theta}) = E_{\theta}[\hat{\theta}(X_1, \dots, X_n)] - \theta$
- Variance: $\text{Var}_{\theta}(\hat{\theta}) = E_{\theta}[(\hat{\theta} - E_{\theta}[\hat{\theta}])^2]$

But are they equally important? How related?

- **Bias-variance decomposition** of square-loss risk

$$E_{\theta}[(\theta - \hat{\theta})^2] = [B(\hat{\theta})]^2 + \text{Var}_{\theta}(\hat{\theta})$$

Extremum estimators

Very general framework that covers elements of major statistical learning frameworks; enjoys good asymptotic behaviour in general!!

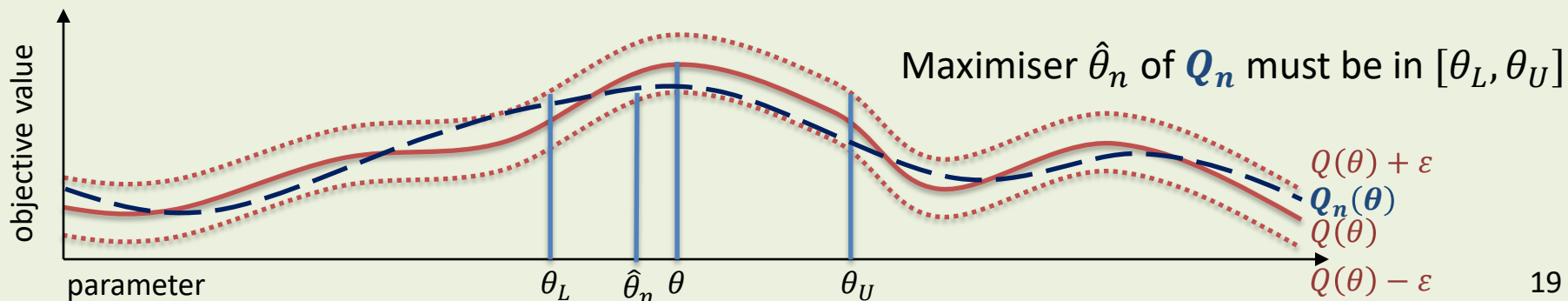
Extremum estimators

- $\hat{\theta}_n(\mathbf{X}) \in \operatorname{argmin}_{\theta \in \Theta} Q_n(\mathbf{X}, \theta)$ for any objective $Q_n()$
- Generalises bits of all statistical frameworks. *Woot!*
 - * **MLE** and **ERM** seen earlier this lecture; and
 - * **MAP** seen later in this lecture.
 - * These are all *M*-estimators, with Q as a sum over data (i.e. of log-likelihood, loss, or log-likelihood plus log prior)
- And it generalises other frameworks too!

Consistency of Extremum Estimators



- Recall consistency: stochastic convergence to 0 bias
- Theorem for extremum estimators: $\hat{\theta}_n \rightarrow \theta$ in prob,
if there's a (“limiting”) function $Q()$ such that:
 - $Q()$ is **uniquely maximised** by θ .
That is, no other parameters make $Q()$ as large as $Q(\theta)$.
 - The parameter family Θ is “**compact**”
(a generalisation of the familiar “closed” & “bounded” set, like $[0,1]$)
 - $Q()$ is a **continuous** function
 - Uniform convergence**: $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \rightarrow 0$ in probability.



A game changer

- Frequentists: estimators that aren't even correct with infinite data (inconsistent), aren't adequate in practice
- Proving consistency for every new estimator? Ouch!
- So many estimators are extremum estimators – general guarantees **make it much easy** (but not easy!) to prove
- **Asymptotic normality**
 - * Extremum estimators converge to Gaussian in distribution
 - * Asymptotic efficiency: the variance of that limiting Gaussian
- Practical: **Confidence intervals** - think error bars!!



→ *Frequentists like to have this asymptotic theory for their algorithms*

Mini Summary

- Decision theory: Utility-based, Minimise risk
- Many familiar learners minimise loss over data (ERM)
- Extremum estimators generalise ERM, MLE, (later: MAP)
 - * Amazingly, consistent: Gives us confidence that they work (eventually)
 - * Amazingly, asymptotically normal: Helps make confidence intervals

Next: Last but not least, the Bayesian paradigm

Bayesian Statistics

Wherein unknown model parameters have associated distributions reflecting prior belief.

Bayesian statistics



Laplace

- Probabilities correspond to **beliefs**
- Parameters
 - * Modeled as r.v.'s having distributions
 - * Prior belief in θ encoded by **prior distribution** $P(\theta)$
 - Parameters are modeled like r.v.'s (even if not really random)
 - Thus: data likelihood $P_{\theta}(X)$ written as conditional $P(X|\theta)$
 - * Rather than point estimate $\hat{\theta}$, Bayesians update belief $P(\theta)$ with observed data to $P(\theta|X)$ the **posterior distribution**

Tools of probabilistic inference

- Bayesian probabilistic inference

- * Start with prior $P(\theta)$ and likelihood $P(X|\theta)$
- * Observe data $X = x$
- * Update prior to posterior $P(\theta|X = x)$



Bayes

- Primary tools to obtain the posterior

- * **Bayes Rule**: reverses order of conditioning

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

- * **Marginalisation**: eliminates unwanted variables

$$P(X = x) = \sum_t P(X = x, \theta = t)$$

This quantity
is called the
evidence

These are
general tools of
probability and
not specific to
Bayesian
stats/ML

Example

- We model $X|\theta$ as $N(\theta, 1)$ with prior $N(0,1)$
- Suppose we observe $X=1$, then update posterior

$$\begin{aligned} P(\theta|X=1) &= \frac{P(X=1|\theta)P(\theta)}{P(X=1)} \\ &\propto P(X=1|\theta)P(\theta) \\ &= \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(1-\theta)^2}{2}\right) \right] \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right) \right] \\ &\propto N(0.5, 0.5) \end{aligned}$$

NB: allowed to push **constants** out front and “ignore” as these get taken care of by normalisation

$$P(\theta|X=1) = \frac{P(X=1|\theta)P(\theta)}{P(X=1)}$$

Name of the game is to get posterior into a recognisable form.
exp of quadratic *must* be a Normal

$$\propto P(X=1|\theta)P(\theta)$$

Discard constants w.r.t θ

$$= \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(1-\theta)^2}{2}\right) \right] \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\theta^2}{2}\right) \right]$$

Collect exp's

$$\propto \exp\left(-\frac{(1-\theta)^2 + \theta^2}{2}\right)$$

$$= \exp\left(-\frac{2\theta^2 - 2\theta + 1}{2}\right)$$

$$= \exp\left(-\frac{\theta^2 - \theta + \frac{1}{2}}{2 \cdot \frac{1}{2}}\right)$$

$$= \exp\left(-\frac{\theta^2 - \theta + \frac{1}{4}}{2 \cdot \frac{1}{2}}\right) \cdot \exp\left(-\frac{\frac{1}{4}}{2 \cdot \frac{1}{2}}\right)$$

$$\propto \exp\left(-\frac{\theta^2 - \theta + \frac{1}{4}}{2 \cdot \frac{1}{2}}\right)$$

Factorise

$$= \exp\left(-\frac{(\theta - \frac{1}{2})^2}{2 \cdot \frac{1}{2}}\right)$$

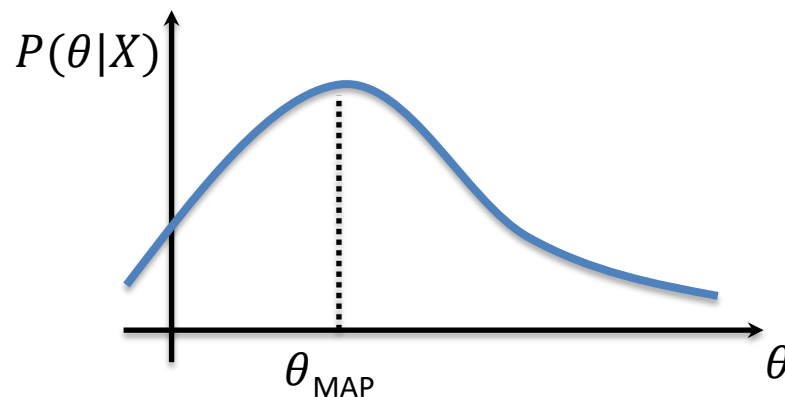
Recognise as (unnormalized) Normal!

$$\propto \mathcal{N}(0.5, 0.5)$$

Constant underlined
Variance/std deviation circled

How Bayesians make point estimates

- They don't, unless forced at gunpoint!
 - * The posterior carries full information, why discard it?
- But, there are common approaches
 - * Posterior mean $E_{\theta|X}[\theta] = \int \theta P(\theta|X) d\theta$
 - * Posterior mode $\operatorname{argmax}_{\theta} P(\theta|X)$ (**max a posteriori** or MAP)
 - * There're Bayesian decision-theoretic interpretations of these



MLE in Bayesian context

- MLE formulation: find parameters that best fit data
$$\hat{\theta} \in \operatorname{argmax}_{\theta} P(X = x|\theta)$$
- Consider the **MAP** under a Bayesian formulation
$$\begin{aligned}\hat{\theta} &\in \operatorname{argmax}_{\theta} P(\theta|X = x) \\ &= \operatorname{argmax}_{\theta} \frac{P(X = x|\theta)P(\theta)}{P(X = x)} \\ &= \operatorname{argmax}_{\theta} P(X = x|\theta)P(\theta)\end{aligned}$$
- **Prior** $P(\theta)$ weights; MLE like *uniform* $P(\theta) \propto 1$
- **Extremum estimator**: $\operatorname{Max} \log P(X = x|\theta) + \log P(\theta)$

Frequentists vs Bayesians – Oh My!

- Two key schools of statistical thinking
 - * Decision theory complements both
- Past: controversy; animosity; almost a 'religious' choice
- Nowadays: deeply connected

I declare the Bayesian vs. Frequentist debate over for data scientists

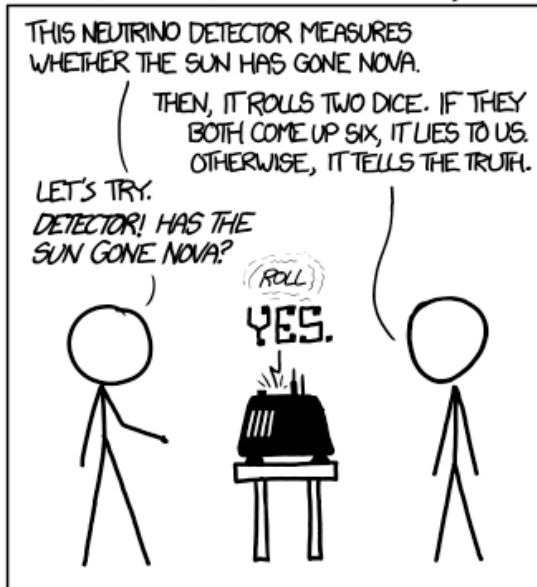
♣ Rafael Irizarry 2014/10/13

Are You a Bayesian or a Frequentist?

Michael I. Jordan
 Department of EECS
 Department of Statistics
 University of California, Berkeley

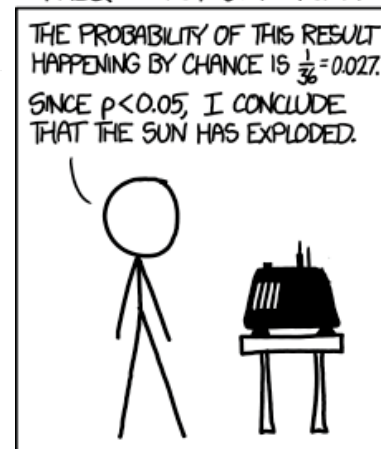
<http://www.cs.berkeley.edu/~jordan>

DID THE SUN JUST EXPLODE?
 (IT'S NIGHT, SO WE'RE NOT SURE.)



<https://xkcd.com/1132/> CC-NC2.5

FREQUENTIST STATISTICIAN:



BAYESIAN STATISTICIAN:



(Some) Categories of Probabilistic Models

Parametric vs non-parametric models

Parametric	Non-Parametric
Determined by fixed, finite number of parameters	Number of parameters grows with data, potentially infinite
Limited flexibility	More flexible
Efficient statistically and computationally	Less efficient

Examples to come! There are non/parametric models in both the frequentist and Bayesian schools.

Generative vs. discriminative models

- X 's are instances, Y 's are labels (supervised setting!)
 - * Given: i.i.d. data $(X_1, Y_1), \dots, (X_n, Y_n)$
 - * Find model that can predict Y of new X
- Generative approach
 - * Model full joint $P(X, Y)$
- Discriminative approach
 - * Model conditional $P(Y|X)$ only
- Both have pro's and con's

Examples to come! There are generative/discriminative models in both the frequentist and Bayesian schools.

Mini Summary

- Bayesian paradigm: Its all in the prior!
- Bayesian point estimate: MAP (an extremum estimator)
- Parametric vs Non-parametric models
- Discriminative vs. Generative models

Next: Logistic regression (unlike you've ever seen before)

Workshops week #2: learning Bayes one coin flip at a time!