1 Summary of the unwritten research.

1. Still on Baehrens

   Doing more reading on Gaussian Processes and overall a review of Markov Chain and other stochastic processes to get a better Idea. This is mostly through Tsitsilkis's Probability and other references or lectures found online.

   Some more Baherens rabbit hole:

   1. obrezanova2009. States that GPs tend to produces more predictive models than their counterparts.
      - GPs seem to be inherently resistant against overfitting(?)
      - It offers the ability to show uncertainty in predictions.
      - Tends to be robust for nonlinear regression.
      - can be applied to both regression and classification.
      - **Methods and Data**:
        - Goal for the GP Gaussian Proess Model: Is to model the probability distribution of the class Y for a molecule, given its descriptor vector $\mathbf{x}, p(y|\mathbf{x})$.
   2. claywood2017 where they talk about using Gaussian Processes for more interpretability of the t tasks at hand.

2. Bayesian Inferenec from good old Tsitsilkis & Bertsekas
   - Bayesian Statistical Inference: This is the process of extracting informaiton about an unknown variable or model by using available data.
   - What makes statistics a bit different from probability: Probability relies on axioms, some assumptions and the consequences of their combinations; probability, on the other hand, can yield different answers to the same question(all of them could be reasonable).
   - Within the field of statistics the are two main schools of though:
     - **Bayesian:** Treats unkown variables as random variables with known (prior) ditributions. This is done by asigning a random variable Θ that caracterizes the model and by postulatinag a **prior probability** distribution $p_\Theta(\theta)$. We then would use *Baye's rule* to derive a **posterior probability** distribution $p_{\Theta|X}(\theta|x)$
     - **Classical/Frequentist:** Unkown variables are treated as *quantities* that happen to be unkown
     - They differ in the way they view of the nature of unkown models or variables

- Two types of inferences:
  - **Model inference**: we studya a real phenomenom or process. Model can then **make predictions of the future**. Model of some planetary trajectory
  - **Variable Inference**: Estimate the value of one or more unkown variables, by using some info. e.g. values send via noisy channel
  - Some blurred lines in this distinction.
- Classification of **Statistical Inference Problems**
  1. **Estimation**: Model is known but we only want to estimate a (possibly multimdimensional) parameter θ(which could be viewed as a random variable).
  2. **Binary Hypthesis Testing** problem: Two hyptothesis exists and data is used to decide which of the two is true. More generally the **m-ary hypthesis testing** for m hypothesis to be tested.
- There are problems were the uncertain object **cannot** be described by a **fixed** number of **parameters**. These are called **nonparametric** problems/models.
- **Inference Methods**:
  1. **Maxium a posterior probability**(MAP): Select the parameter/ hypothesis withmaximum conditional/posterior probability given the data.
  2. **Least Mean Squares**(LMS) estimation: Selet estimator/model that *minimizes* the mean squared error between parameter and estimate.
  3. **LInear least mean squares**: Select an estimator which is a linear funciton of the daa and minimizes the mean squared error between the parameter and its estimate.
- **Bayesian Inference and the Posterior Distribution**
  - Assume we know:
    1. The joint distribution of Θ and X
    2. Prior distriution p_Θ(discrete) or f_Θ(continuous)
    3. Conditional DIstribution $p_{X|\Theta}$
  - Once we get a sample x from X we can use posterior distribution $p_{\Theta|X}$ of Θ to solve Bayesian inference problem.

2 Gaussian Process Regression From First Principles

article

Found gold baby. The paper linked is a very quick and concise intro to gaussian processes for regression.

One way to view Gaussian Processes is as a distribution *over functions*. Given the paremeters that define the GP(Mean and covariance matrices) we can sample a function at the point $\mathbf{x} \in R^d$ (R for reals)

The sampling goes as follows:

$f(\mathbf{x})\ GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$

Where m() is for mean and k(,) for covariance

So it seems to be that if we have our set

$\mathbf{X}_i, \ldots, \mathbf{X}_j\ N(\mu, \Sigma)$

Then those random variables are used to sample $f(\mathbf{X})$ on every step. Remember that we read that funcitons can sort of be thought as mapping one vector to another. Maybe we are mapping independent random variables to their samples and that forms a function. When we sample from the entire set once again then the new vector of outputs that we get will yield a different function.

No, no. I think that its rather that each random variable can yield a function.

GIven an a dataset, GPR predicts a poserterior Gaussian Distribution for targets over test points. BY computing the parameters of this Gaussian distribution given observed training data.

This *non-parametric* property of GPR seems to bee all the rage

3 Some explanation of the Kernel Trick

Source

Goal woud lbe to linearly separate previously unseparable elements in higher dimensions. Best example in article is $\theta(x) = x \mod 2$

Basically its the art of selecting the right transformation to make some boundary that is a n-1 hyperplane of the extended n dimensions.

The support vectors are the vectors that describe the points that allow us to specify a hyperplane that defines our decision boundary.

3.1 The actual explanation now

SO whats the problem with this ? It seems that this higher dimensional computations would naturally become expensive and prohibitive. The **Kernel Trick** solves this by using methods that represent our data using a set of pairwise similarity comparisons between the orignal data observations instal of explicitly applying the transformations and representing the data by these transformed coordiantes in higher dimensional feature space.

4 Limit Theorems

These theorems are related to the asymptotic behavior of sequence of random variables

Im not sure if im understanding this kernel trick .