ANLY601 Assignment2

Shaoyu Feng (sf865)
Collobrator: Mengtong Zhang
Collobrator: Yunjia Zeng

Feb 2020

1 Question 1: Convexity

1. $f(x) = \sum_{i=1}^{\infty} ||x||_p, p > 0$
   Ans:
   It is convex.
   Based on the Triangle inequality of norm, we have $||v + w|| \leq ||v|| + |w|||$ for any $p > 0$. Based on the definition of convexity, we can see any form norm is convex, since we have

   $$||\lambda v + (1-\lambda)w|| \leq ||\lambda v|| + ||(1-\lambda)w|| = \lambda||v|| + (1-\lambda)||w||$$

   Since each norm form is convex, the sum of norm is convex as well.

2. $f(d) = k(x, x') - k'(x, x')$ where k(x,x’)=k(x-x’)=k(d) is also a stationary and positive kernel.
   Ans:
   It is not necessary convex. For example, Assume we have $k(d) = k(x, x') = d$ and $k(d) = k(x, x') = d^2$. This is clearly not convex, we have $f(x) = d - d^2$, this clearly does not satisfy the definition of convexity. For example, let $\lambda = 0.5, v = 0, w = 1$ we have $f(\lambda v + (1-\lambda) = f(0.5) = 0.25 > f(\lambda v) + f((1-\lambda)w) = 0$.

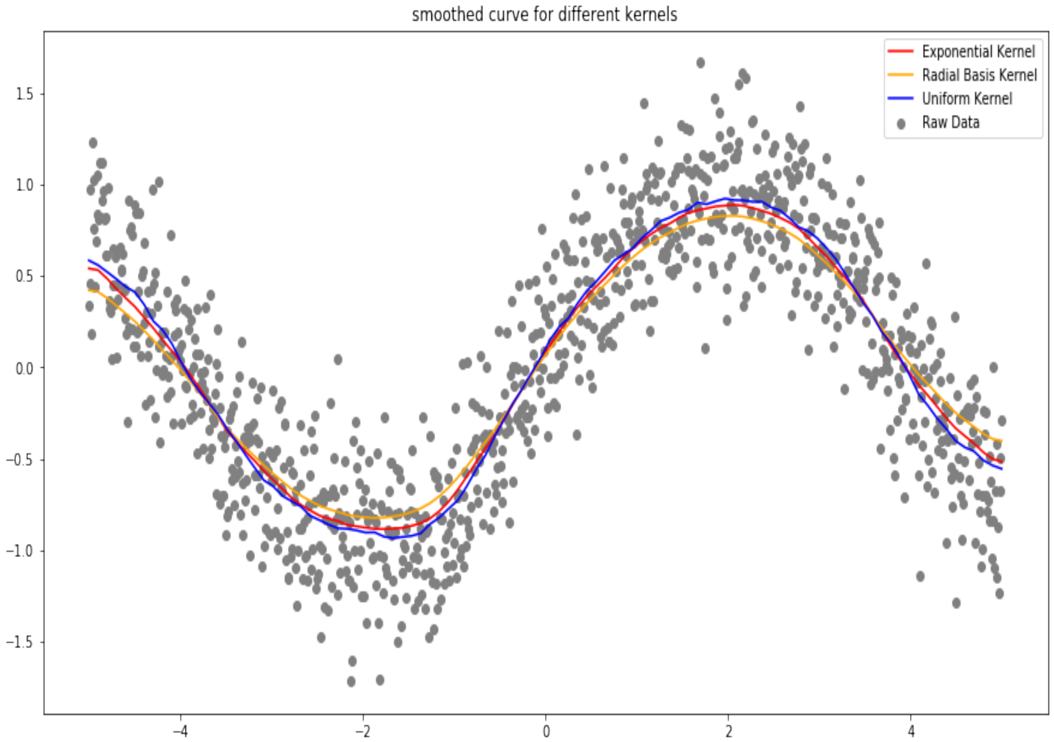
3. $f(d) = k(x, x') * k'(x, x') - b$ for some $b \in R$
   Ans:
   Not necessarily convex. Assume we have $k(d) = k(x, x') = d$ and $k(d) = k(x, x') = -d$. we then have $f(x) = -d^2 - b$ This is not convex, counter-example not obeying definition of convexity can be easily found.

4. $f(d) = |x||_p - max(0, x)$
   Ans:
   it is convex.
   we have $f(d) = |x||_p - max(0, x) = |x||_p + min(0, -x)$.
   $\|\lambda v + (1-\lambda)w\|_p + min(0, -\lambda v - (1-\lambda)w) \leq \|\lambda v\|_p + \|(1-\lambda)w\|_p + min(0, -\lambda v) + min(0, -(1-\lambda)w) = \lambda(\|v\|_p + min(0, -v)) + (1-\lambda)(\|w\|_p + min(0, -w))$
   From the definition of convexity, $f(x)$ is convex.

5. $f(d) = |x||_p + max(0, x)$ for some $b \in R$
   Ans:
   Yes, it is convex.
   we can prove that max(0,x) is convex since $max(0, \lambda v + (1-\lambda)w) \leq \lambda max(0, x) + (1-\lambda)max(0, x)$
   Any positive combination of convex function is convex.

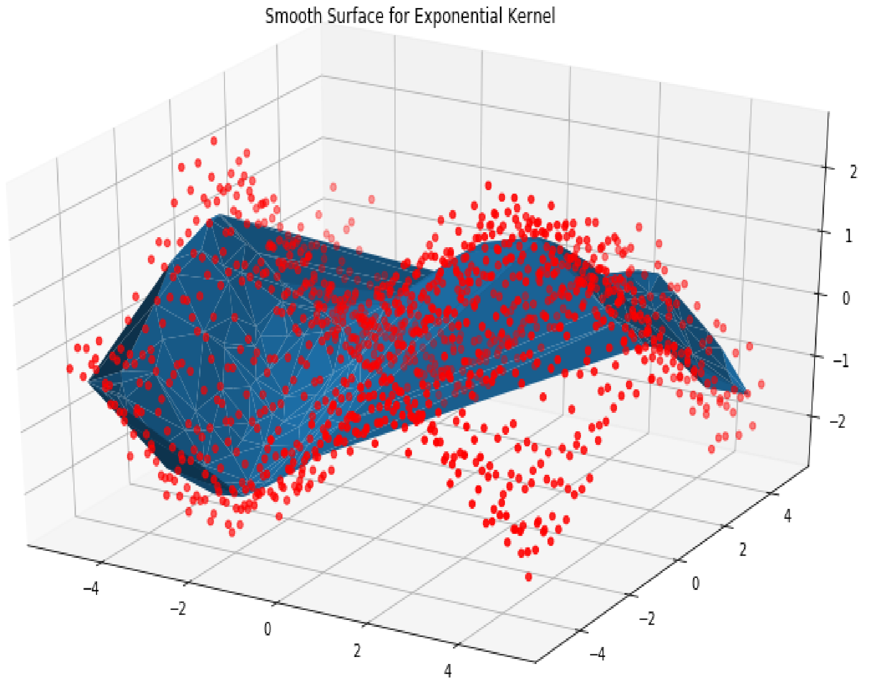
2 Question 2: Kernel Regression

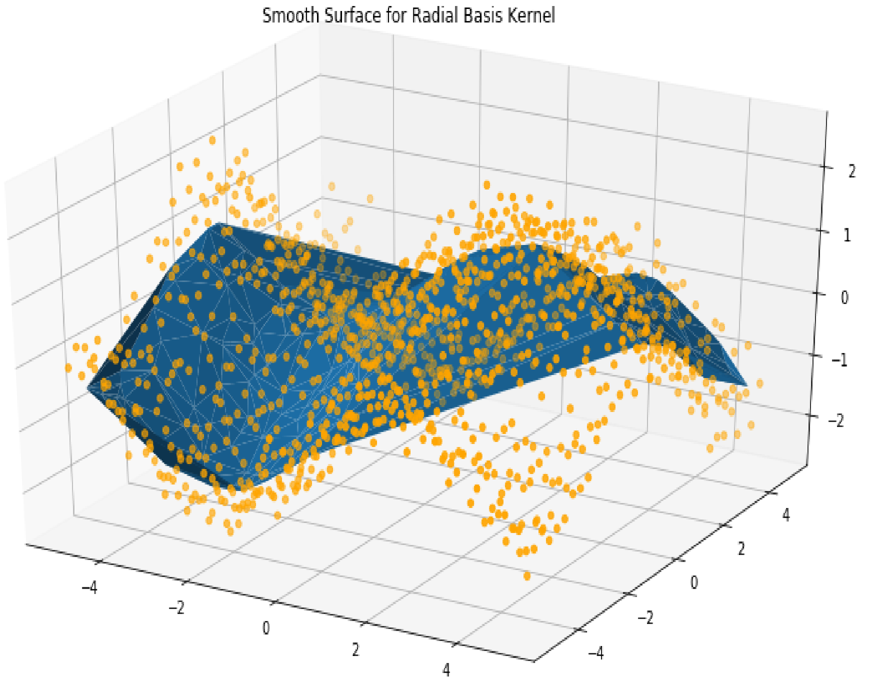
Code: See Jupyter Notebook
Part1:

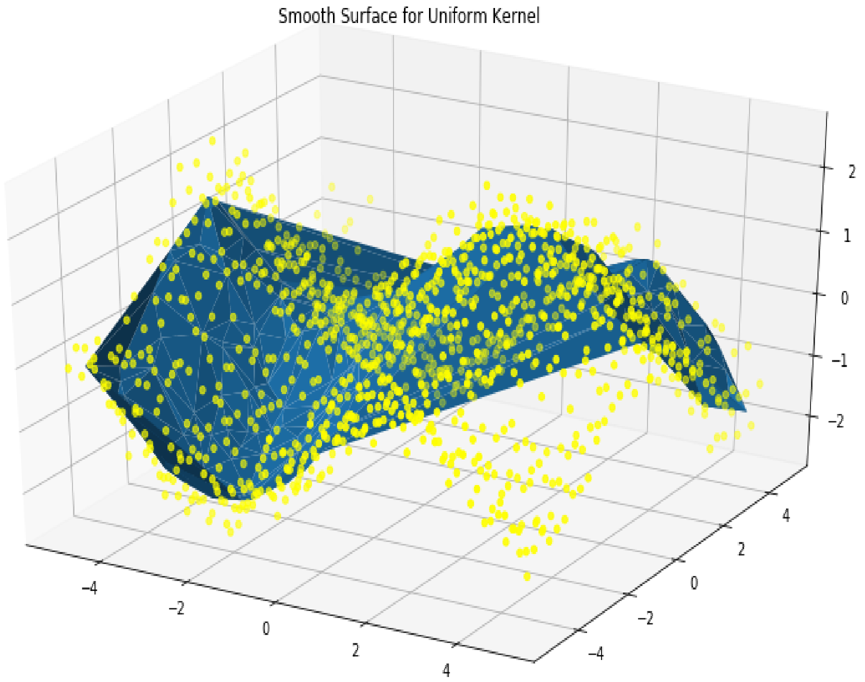


Part2:
As we observe from above plot, we can see all three kernels provides a good fit for the data. Radial and Exponential are comparably more smooth than Uniform Kernel,
Part3::

we can see that the kernels still fit the data reasonably well, despite some areas around x=4 and y in -2 to -4.

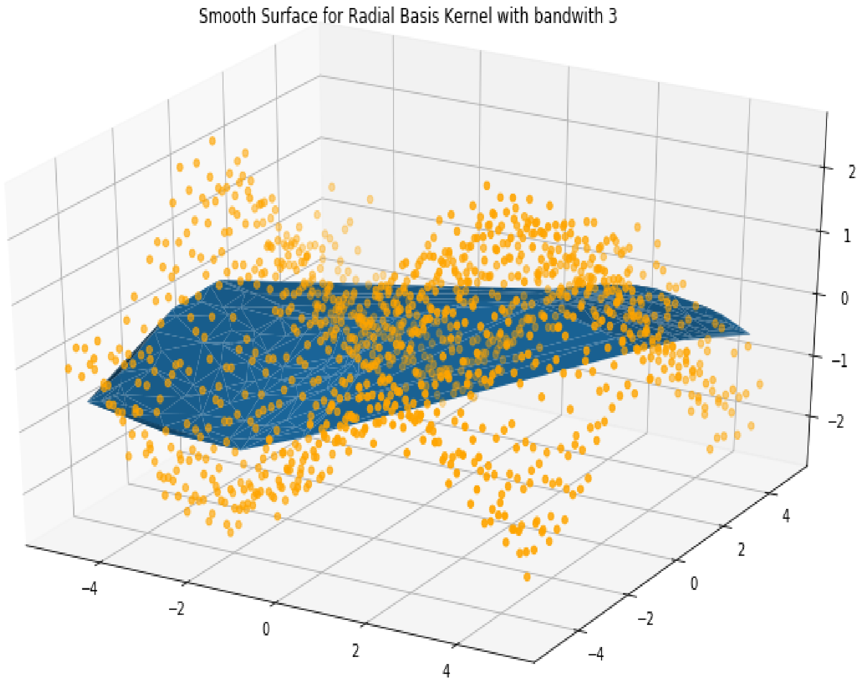

Smooth Surface for Exponential Kernel
2
1
0
−1
−2
−4
−2
0
2
4
4
2
0
−2
−4

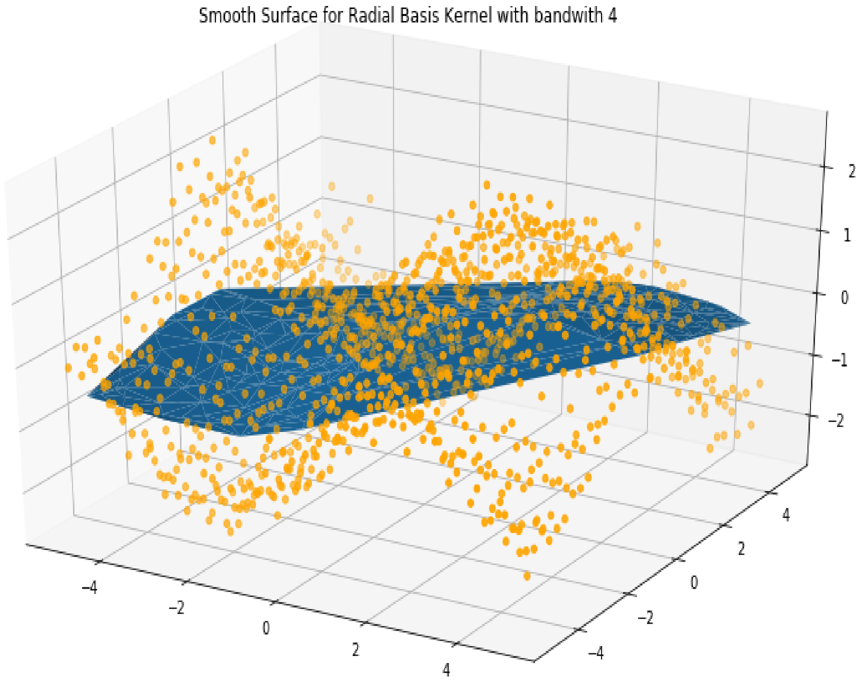

Smooth Surface for Radial Basis Kernel
2
1
0
−1
−2
−4
−2
0
2
4
4
2
0
−2
−4

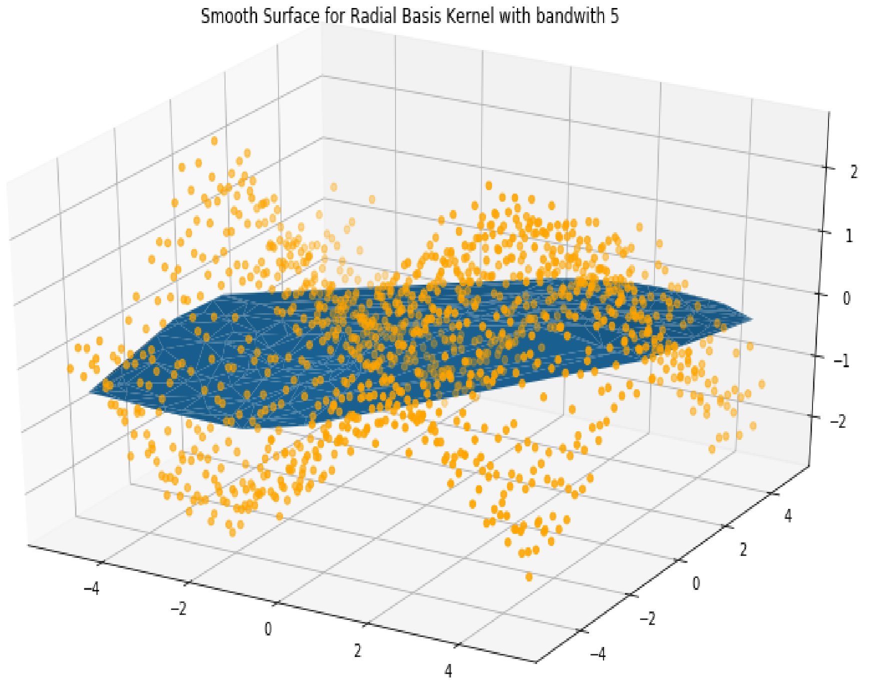


After Change Bandwidth::

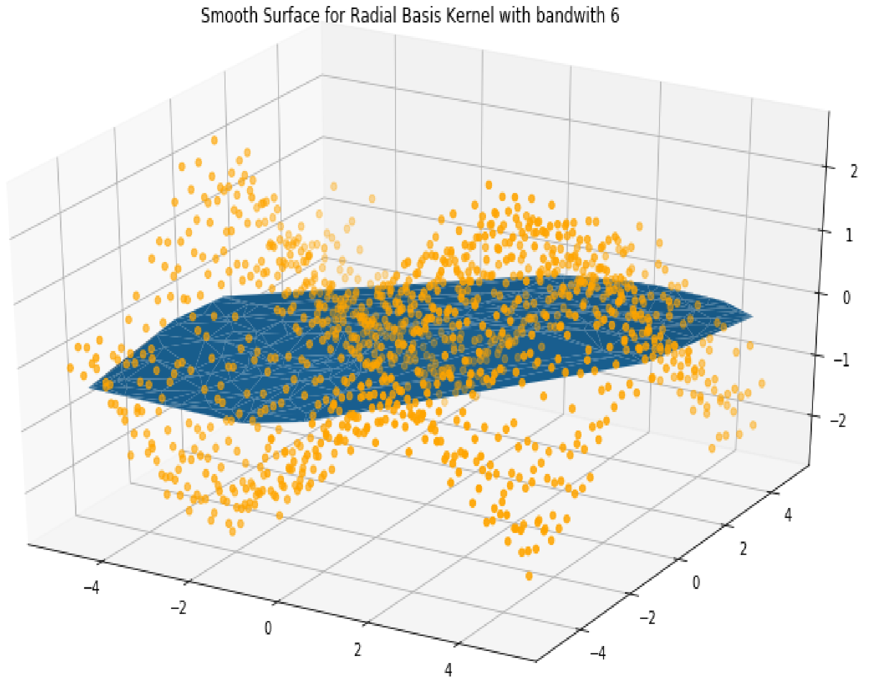
As we observe below, we notice that we increase the bandwidth, the fitted plane becomes more and more flat.


Smooth Surface for Radial Basis Kernel with bandwith 3
−4
−2
0
2
4
−4
−2
0
2
4
−2
−1
0
1
2


Smooth Surface for Radial Basis Kernel with bandwith 4
−4
−2
0
2
4
−4
−2
0
2
4
−2
−1
0
1
2

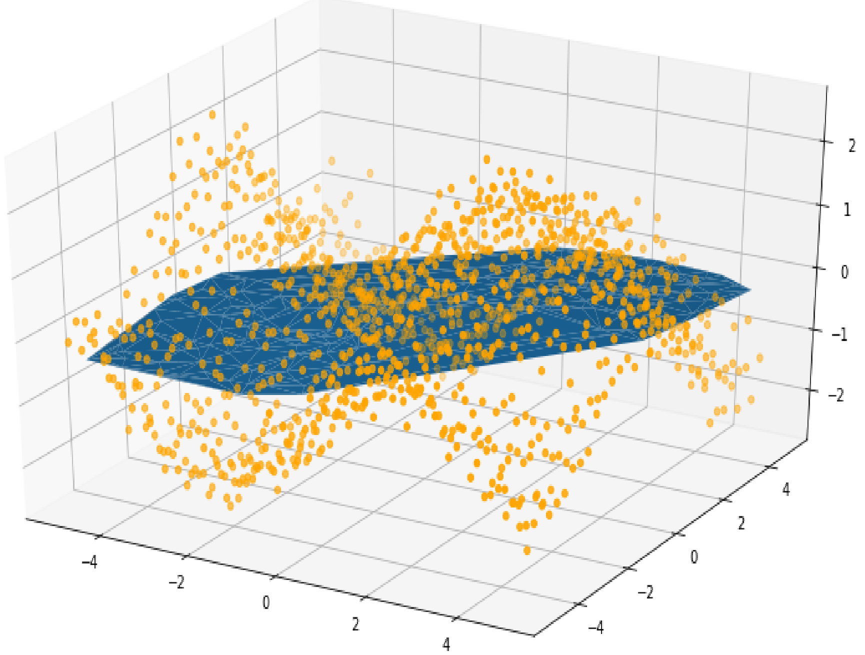


Smooth Surface for Radial Basis Kernel with bandwith 5

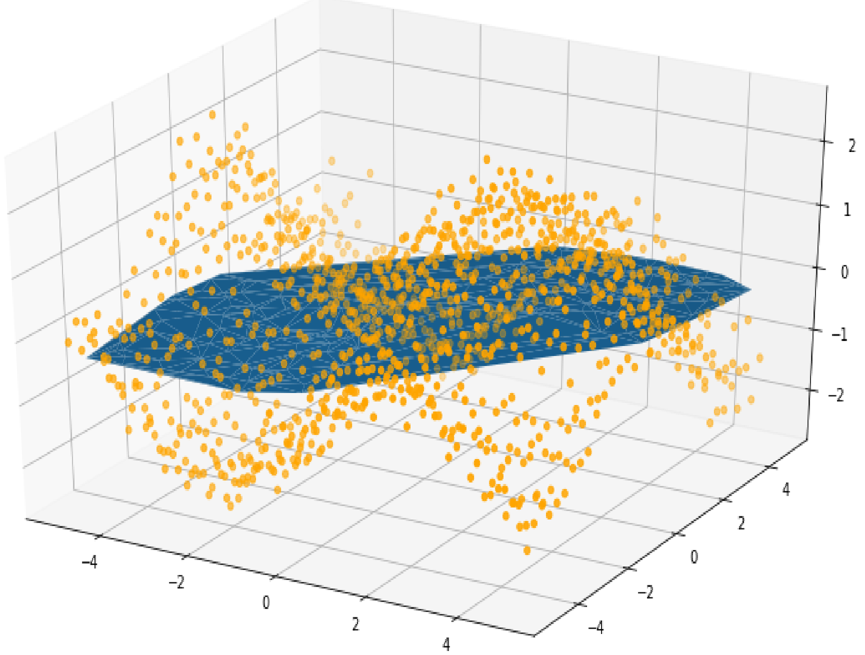


Smooth Surface for Radial Basis Kernel with bandwith 6

Smooth Surface for Radial Basis Kernel with bandwith 7



Smooth Surface for Radial Basis Kernel with bandwith 8

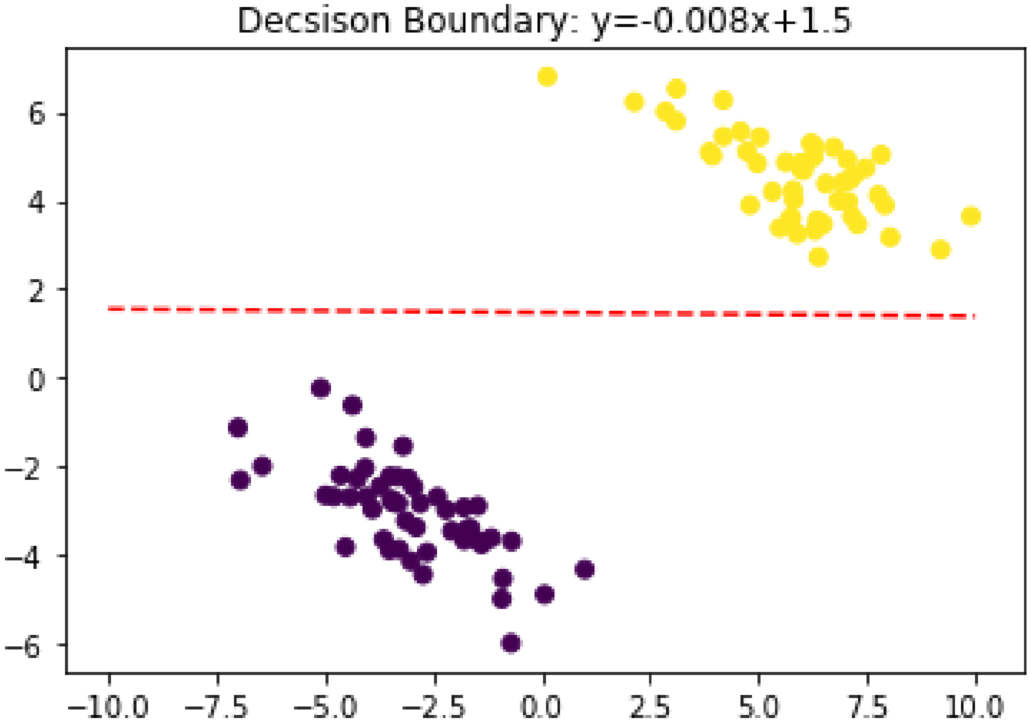


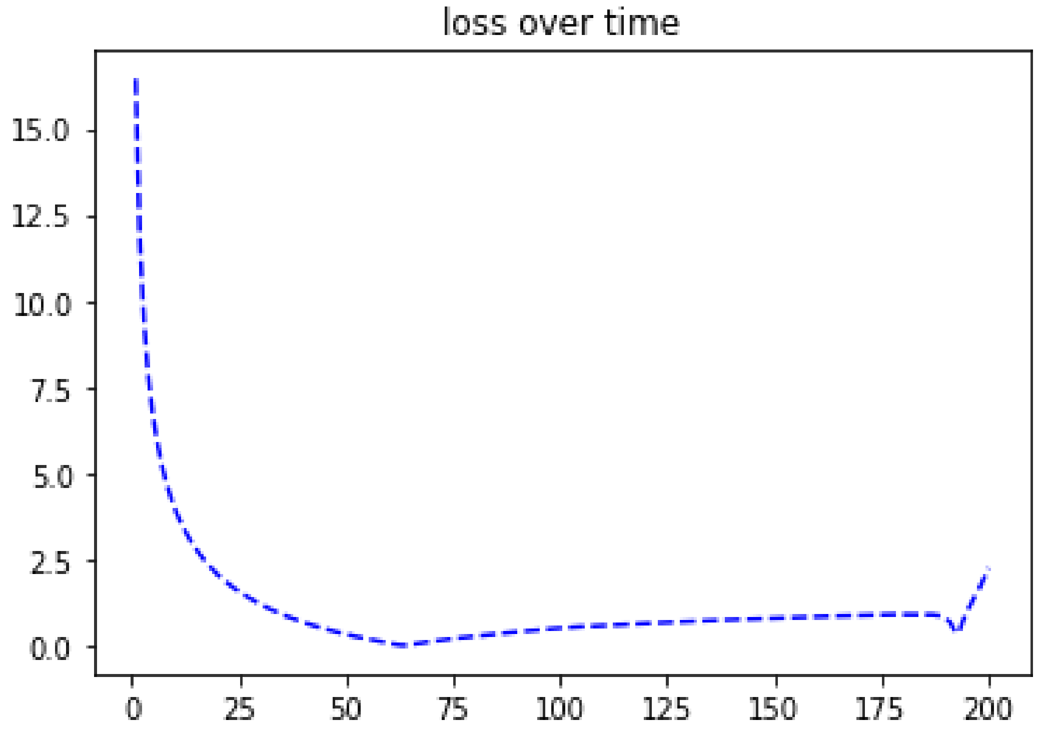
3 Question 3: Programming: Stochastic Subgradient Descent

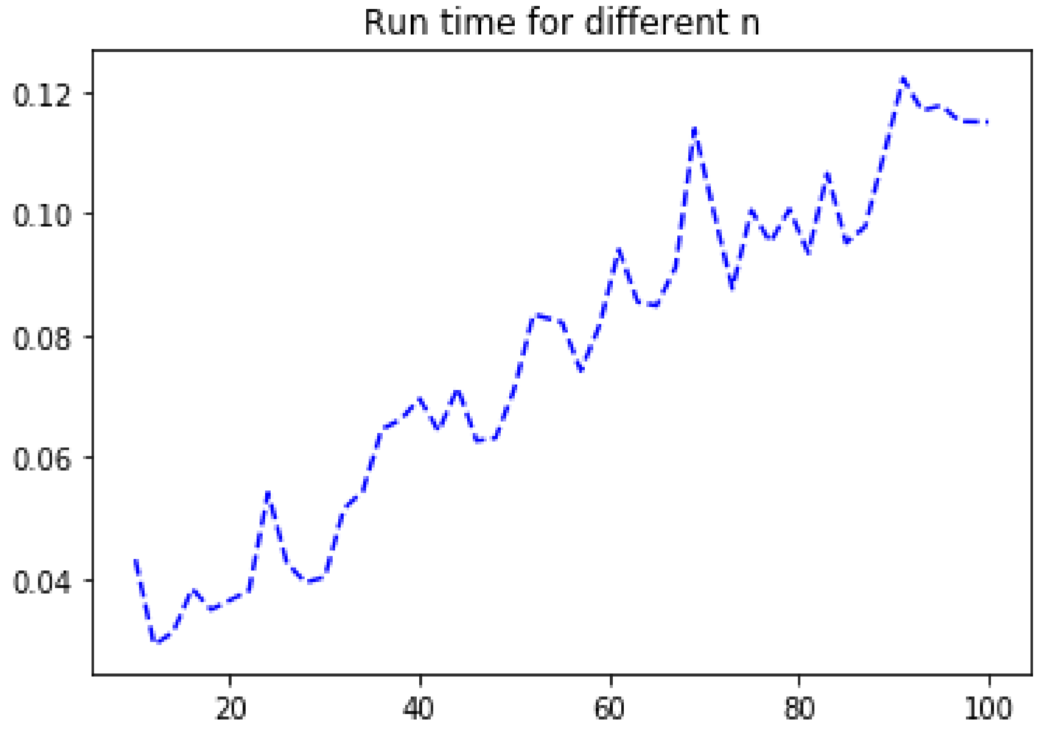
Code: See Jupyter Notebook
Write-up:

To speed-up the run-time, instead of using all data set during each epoch, i am random-sampling a subset of data set to perform the training. The algorithm is able to converge given that the number of epoch is big enough(¿20).

Following three graphs show the decision boundary, hinge loss over time and run time as n increases.







As we can see, the decision boundary is able to separate the two class well, however, it is not max-margin decision boundary. The loss shows that algorithm is able to converge. Last chart shows that the run-time is almost linear but not entirely.

4 Question 4: Calculating the conjugate distribution

1. $\mu \sim N(\tau, v)$, $\sigma^2 \sim InverseGamma(\alpha, \beta)$ where, $X \sim N(\mu, \sigma^2)$ and τ, v, α, are all constant.
   Ans:
   Firstly, for conjugate posteriors for μ we have:

$$P(\mu|\tau, \nu, \sigma, \mathbf{x}) \propto \frac{\mathbf{1}}{\sqrt{\mathbf{2}\pi\nu}}\mathbf{e}^{-\frac{(\mu-\tau)^{\mathbf{2}}}{\mathbf{2}\nu}} * \prod_{\mathbf{i=1}}^{\mathbf{n}} \frac{\mathbf{1}}{\sqrt{\mathbf{2}\pi\sigma}}\mathbf{e}^{-\frac{(\mathbf{x_i}-\mu)^{\mathbf{2}}}{\mathbf{2}\sigma}}$$

$$P(\mu|\tau, \nu, \sigma, \mathbf{x}) \propto \mathbf{constant} * \mathbf{e}^{\frac{(\mu-(\tau\sigma+\mathbf{n\bar{x}v})/(\sigma+\mathbf{n}\nu))^{\mathbf{2}}}{\mathbf{2}\nu\sigma/(\sigma+\mathbf{n}\nu)}}$$

   From above, we can see that:

$$\mu|\tau, \nu, \sigma, \mathbf{x} \sim \mathbf{Normal}(\frac{\tau\sigma + \mathbf{n\bar{x}v}}{\sigma + \mathbf{n}\nu}, \frac{\nu\sigma}{\sigma + \mathbf{n}\nu})$$

   For conjugate posteriors for σ^2 we have:

$$P(\sigma|\alpha, \beta, \mu, \mathbf{x}) \propto \frac{\beta^\alpha}{\mathbf{\Gamma}(\alpha)}(\frac{\mathbf{1}}{\sigma^{\mathbf{2}}})^{\alpha+\mathbf{1}}\mathbf{e}^{-\beta/\sigma^{\mathbf{2}}} \prod_{\mathbf{i=1}}^{\mathbf{n}} \frac{\mathbf{1}}{\sqrt{\mathbf{2}\pi\sigma}}\mathbf{e}^{-\frac{(\mathbf{x_i}-\mu)^{\mathbf{2}}}{\mathbf{2}\sigma^{\mathbf{2}}}}$$

$$P(\sigma|\alpha, \beta, \mu, \mathbf{x}) \propto (\frac{\mathbf{1}}{\sigma})^{\alpha+\mathbf{1}+\mathbf{n}/\mathbf{2}}\mathbf{e}^{(\mathbf{2}\beta+\mathbf{\Sigma_{i=1}^n}(\mathbf{x_i}-\mu)^{\mathbf{2}})/\mathbf{2}\sigma^{\mathbf{2}}}$$

   From above, we can see that:

$$\sigma|\alpha, \beta, \mu, \mathbf{x} \sim \mathbf{InverseGamma}(\alpha + \mathbf{n}/\mathbf{2}, \beta + \mathbf{\Sigma_{i=1}^n}(\mathbf{x_i} - \mu)^{\mathbf{2}}/\mathbf{2})$$

2. $p \sim Dirichlet(\alpha_1, ..., \alpha_k)$ where $X \sim Multinoial(p)$ where $p = (p_1, ..., p_k)$, this is an extension of beta binomial.
   Ans:

$$P(p_1, p_2, ..., p_k|x_{ij}) \propto \prod_{i=1}^{n} p_i^{\alpha_i - 1} \prod_{i=1}^{n} \frac{n!}{x_{i1}!...x_{ik}!} p_1^{x_{i1}} ... p_k^{x_{ik}}$$

$$P(p_1, p_2, ..., p_k|x_{ij}) \propto \prod_{i=1}^{n} p_i^{\alpha_i + \Sigma_{j=1}^{n} x_{ji} - 1}$$

   From above, we can see that:

$$p_1, p_2, ...p_k|x_{ij}) \sim Dirichlet(\alpha_1 + \Sigma_{j=1}^{n} x_{j1} - 1, ..., \alpha_k + \Sigma_{j=1}^{n} x_{jk} - 1)$$

3. $\lambda \sim Gamma(\alpha, \beta) where X \sim Possion(\lambda)$
   Ans:

$$P(\lambda|x_1, ..., x_n) \propto \lambda^{\alpha-1}e^{-\lambda/\beta} \prod_{i=1}^{n} \lambda^{x_i} e^{-\lambda} \propto \lambda^{n\bar{x}+\alpha-1}e^{\lambda(-n-1/\beta)}$$

$$\lambda|x_1, ..., x_n \sim Gamma(n\bar{x} + \alpha, \frac{\beta}{\beta n + 1})$$

5 Question 5: Priors as regularizers

Show that in logistic regression:
– The L2 penalty (ridge) is equivalent to a Normal prior
Ans:
Suppose we are estimating β with prior distribution of β as $N(0, \lambda^{-1})$,

$$\hat{\beta}_{MAP} = \arg\max_{\beta} P(\beta|y) = \arg\max_{\beta} \frac{P(y|\beta)P(\beta)}{P(y)} = \arg\max_{\beta} P(y|\beta)P(\beta)$$

In this case, we have

$$\prod_{n=1}^{N} N(y_n|\beta x_n, \sigma^2) * N(\beta|0, \lambda^{-1})$$

Take the log, we have:

$$\sum_{n=1}^{N} -\frac{1}{\sigma^2}(y_n - \beta x_n)^2 - \lambda\beta^2 + const$$

From above, we can see the target function of maximum posterior estimation is equivalent to ridge regression. Thus, The L2 penalty (ridge) is equivalent to a Normal prior.

– The L1 penalty is a LaPlace priors
Ans:
Similarly:

$$\arg\max_{\beta} \Big[\log\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-(\beta_0+\beta_1 x_{i,1}+...+\beta_p x_{i,p}))^2}{2\sigma^2}} + \log\prod_{j=0}^{p} \frac{1}{2b} e^{-\frac{|\beta_j|}{2b}}\Big]$$

$$= \arg\max_{\beta} \Big[-\sum_{i=1}^{n} \frac{(y_i-(\beta_0+\beta_1 x_{i,1}+...+\beta_p x_{i,p}))^2}{2\sigma^2} - \sum_{j=0}^{p} \frac{|\beta_j|}{2b}\Big]$$

$$= \arg\min_{\beta} \frac{1}{2\sigma^2}\big[\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_{i,1}+...+\beta_p x_{i,p}))^2 + \frac{\sigma^2}{b}\sum_{j=0}^{p}|\beta_j|\big]$$

$$= \arg\min_{\beta} \big[\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_{i,1}+...+\beta_p x_{i,p}))^2 + \lambda\sum_{j=0}^{p}|\beta_j|\big]$$

From above, we can see the target function of maximum posterior estimation is equivalent to LASSO regression. Thus, The L1 penalty (LASSO) is equivalent to a Laplace prior.

6 Question 6: General Questions

1. Explain in your own words the difference between the posterior distribution and posterior predictive distribution.

   Ans:

   The simple difference between the two is that the posterior distribution depends on the unknown parameter θ. While on the other hand, the posterior predictive distribution does not depend on the unknown parameter θ because it has been integrated out. And the posterior distribution refers to the distribution of the parameter, while the predictive posterior distribution (PPD) refers to the distribution of future observations of data.

2. Which one would you use to predict future values of X? Explain your rationale.

   Ans:
   I will use posterior predictive distribution. It is the distribution for future predicted data based on the data you have already seen. So the posterior predictive distribution is basically used to predict new data values.

3. show that as n increase for a $X \ N(\mu, \sigma^2)$ where $\mu \sim N(\alpha, \beta)$ and $\sigma^2 \sim InverseGamma(\tau, b)$ that $\mu_{MAP} - > \mu_{MLE}$ and $\sigma^2_{MAP} - > \sigma^2_{MLE}$

   Ans:
   First of all the MLE for μ and σ^2 is given by

$$f(x_1, x_2, \dots, x_n|\sigma^2, \mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = (\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\log(f(x_1, x_2, \dots, x_n|\sigma, \mu)) = n \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Firstly take derivetive of μ,

$$\frac{d\mathcal{L}}{d\mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \ |_{\mu} = 0$$

$$\hat{\mu}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Then we take derivative with regard to σ

$$\frac{d\mathcal{L}}{d\sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n}(x_i - \mu)^2\sigma^{-3} = 0$$

$$\hat{\sigma}^2_{MLE} = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n} = \sigma_x^2$$

Based on the calculation of problem 4 part 1:

$$\hat{\mu}_{MAP} = \frac{\alpha\sigma + n\bar{x}\beta}{\sigma + n\beta}$$

As n increase, MAP is approching $\bar{x}$, which is similar to MLE for μ
Similarly:

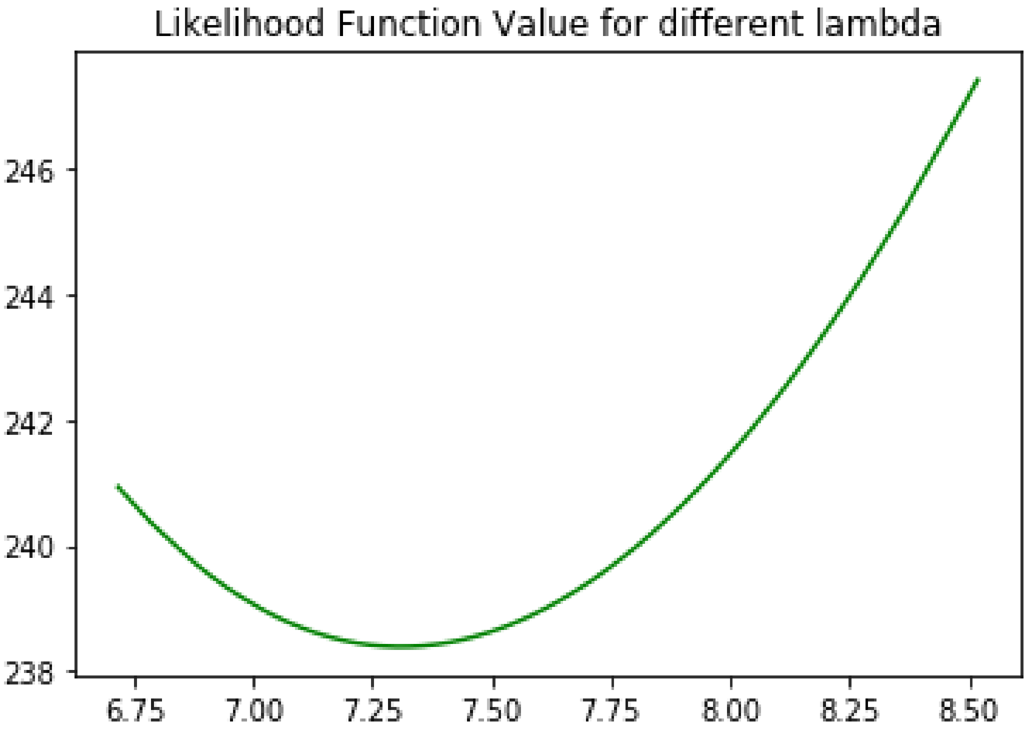
$$\hat{\sigma^2}_{MAP} = \frac{\nu + \Sigma_{i=1}^{n}(x_i - \mu)^2/2}{\tau + n/2 - 1}$$

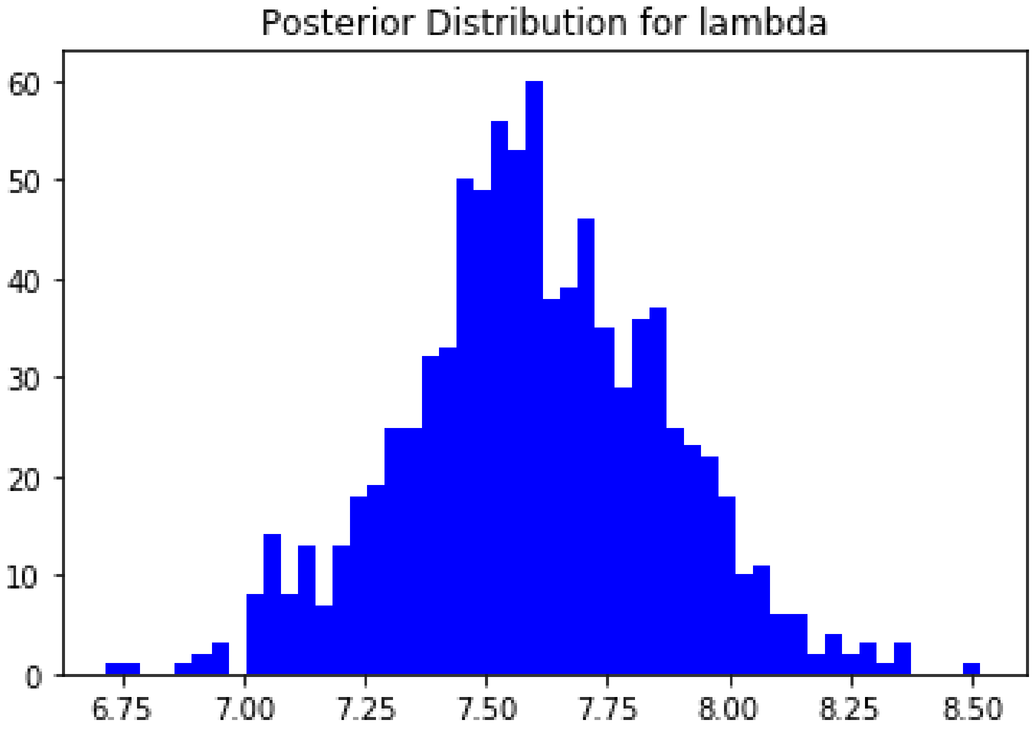
when n is large, it approached to $\frac{\sum_{i=1}^{n} x_i}{n}$, which is similar to MLE results for σ^2

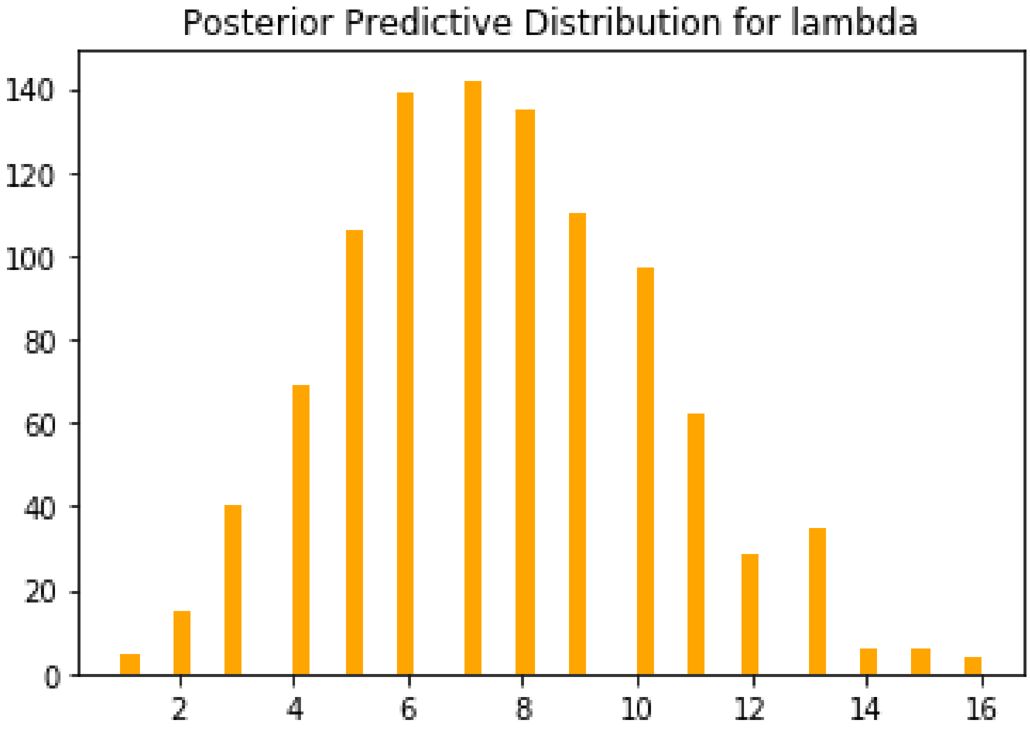
7 Question 7: Programming a Gibbs Sampler

Code: See Jupyter Notebook
Write-up:
Following three charts shows likelihood function values for different lambda, posterior distribution of lambda and posterior predictive distribution. Given for this problem, there is no dependencies of

lambda, there is no convergence problem for lambda.

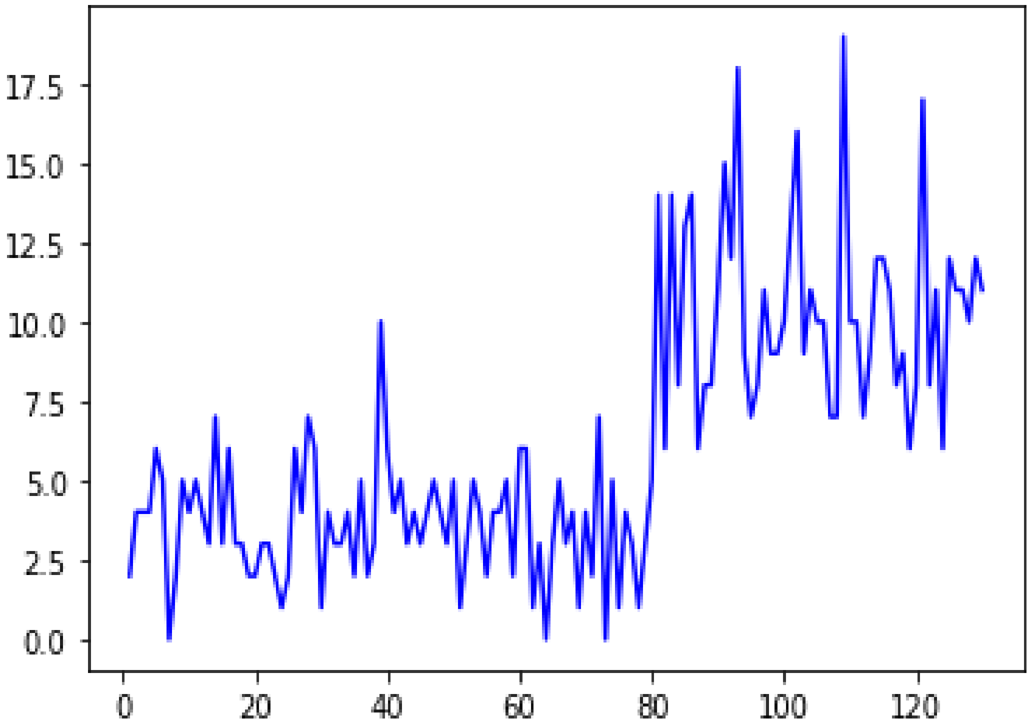






8 Question 8: Change points models

Code: See Jupyter Notebook
Write-up:



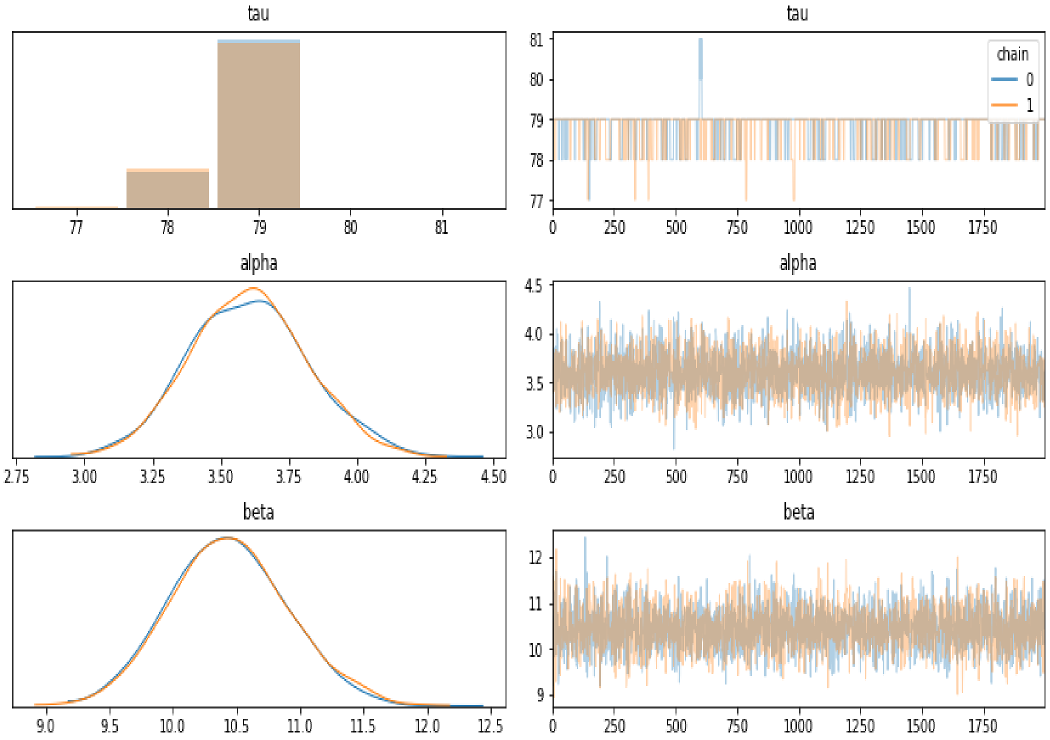
we have $X|\lambda_t \sim (\lambda_t)$ and $lambda \sim exponential(\tau_i)$ where τ_i has different values after switch point. Before change point, we have: $\lambda_t \sim exponential(\alpha)$ we then have:

$$f(\alpha|x) \propto const * e^{-\tau\lambda}\lambda^{\sum_{i=1}^{\tau} x_i} * e^{-\alpha\lambda}$$

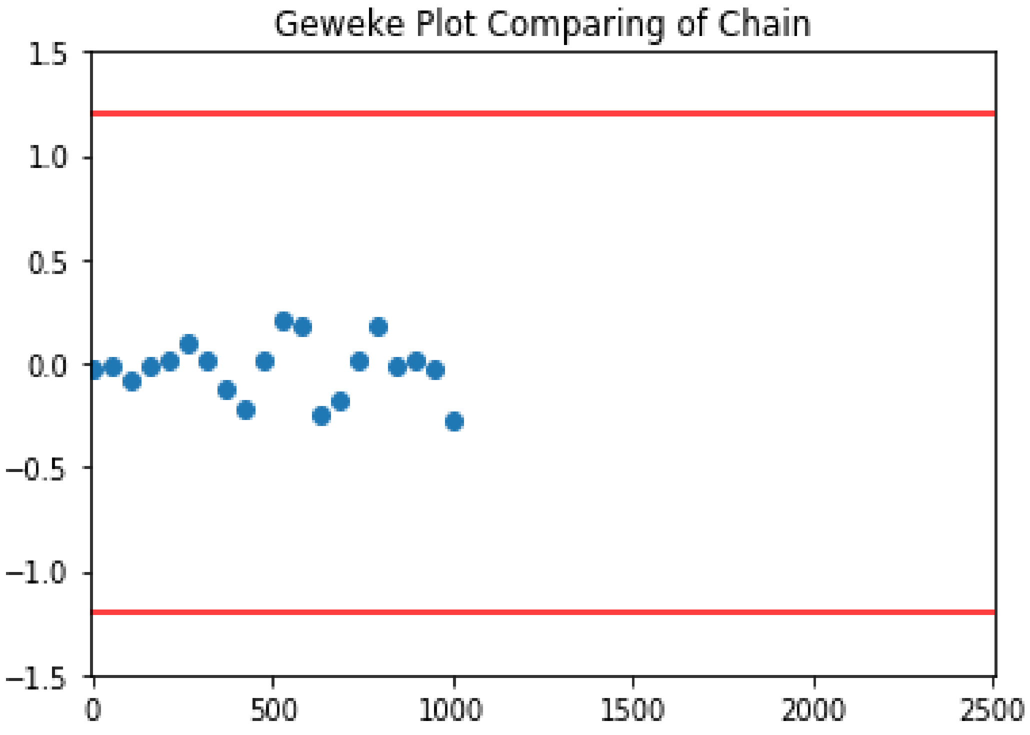
$$f(\alpha|x) \propto \lambda^{\sum_{i=1}^{\tau} x_i} e^{-(\tau+\alpha)\lambda}$$

It follows a Gamma Distribution $(\sum x_i + 1, \alpha + \tau)$. But we dont have to model the paramater estimation in this way.

Using the Bayesian approach, we can estimate all α, β, τ through MCMC, following is the result:



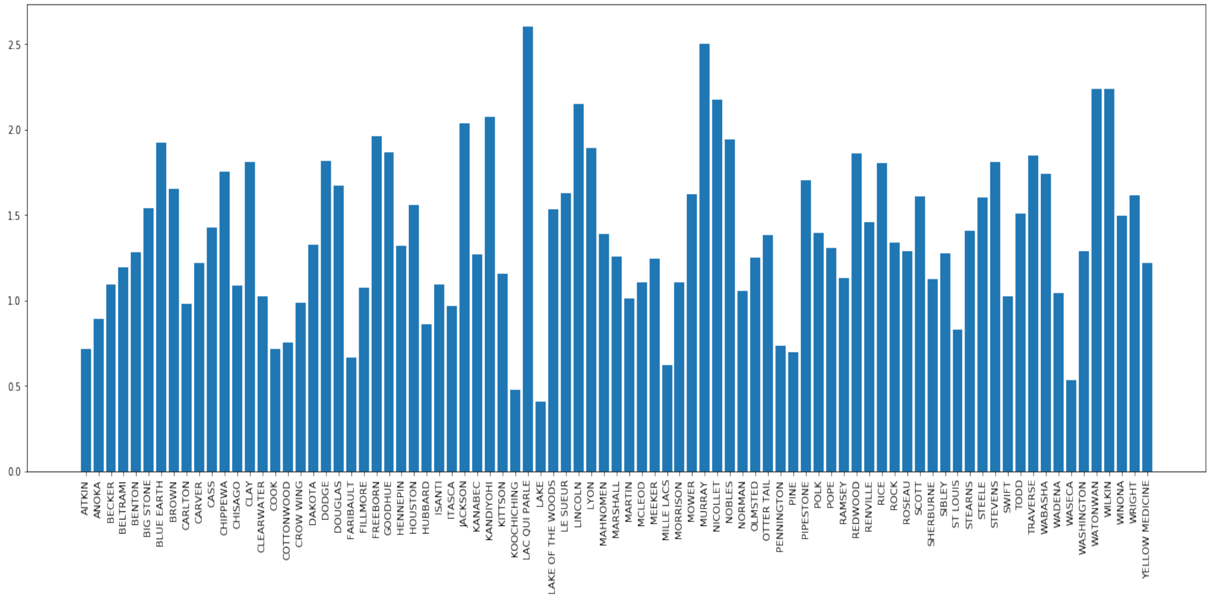
From above, we observe that the τ was mostly identified as 79 through the histogram. and multiple-chains results showed similar results, which suggests that the sampler is converged.



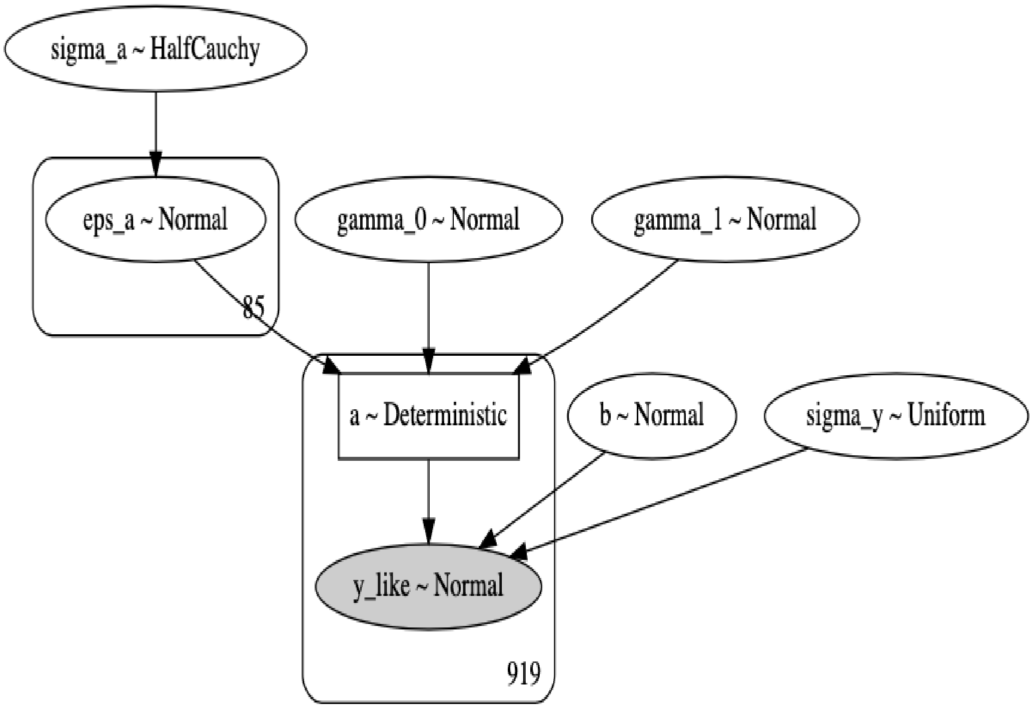
Another way to prove the convergence is through Geweke Plot, we can see that the test value is stable and have a small value, which suggests that the sampler is converged from another angle.

9 Question 9: Programming a hierarchical model using PYMC3

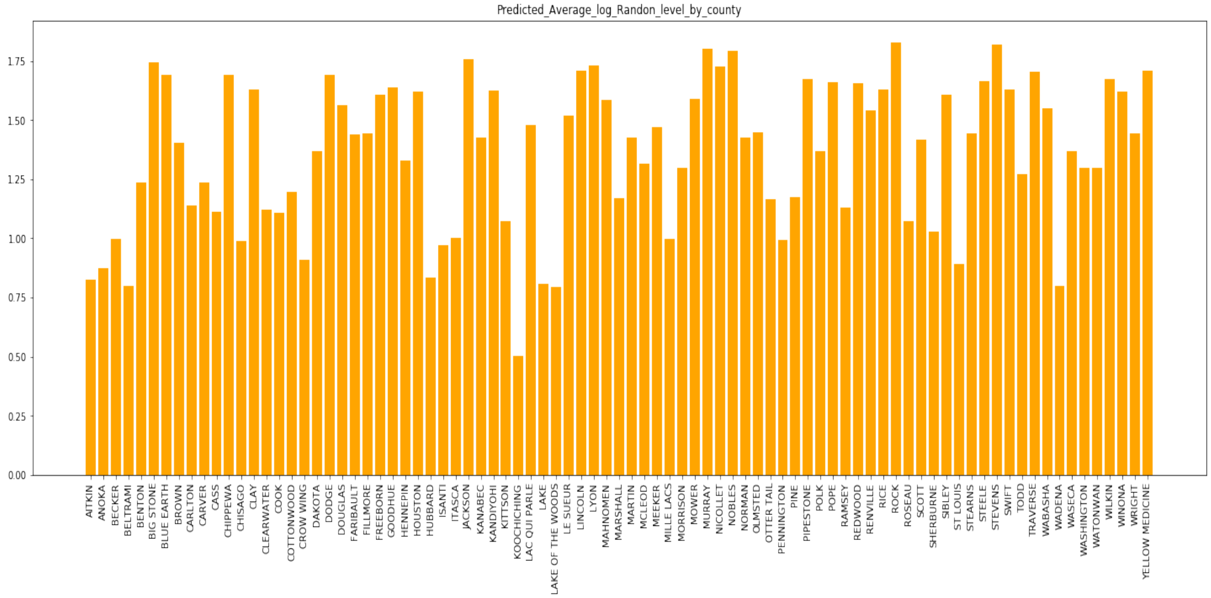
Code: See Jupyter Notebook
Write-up:



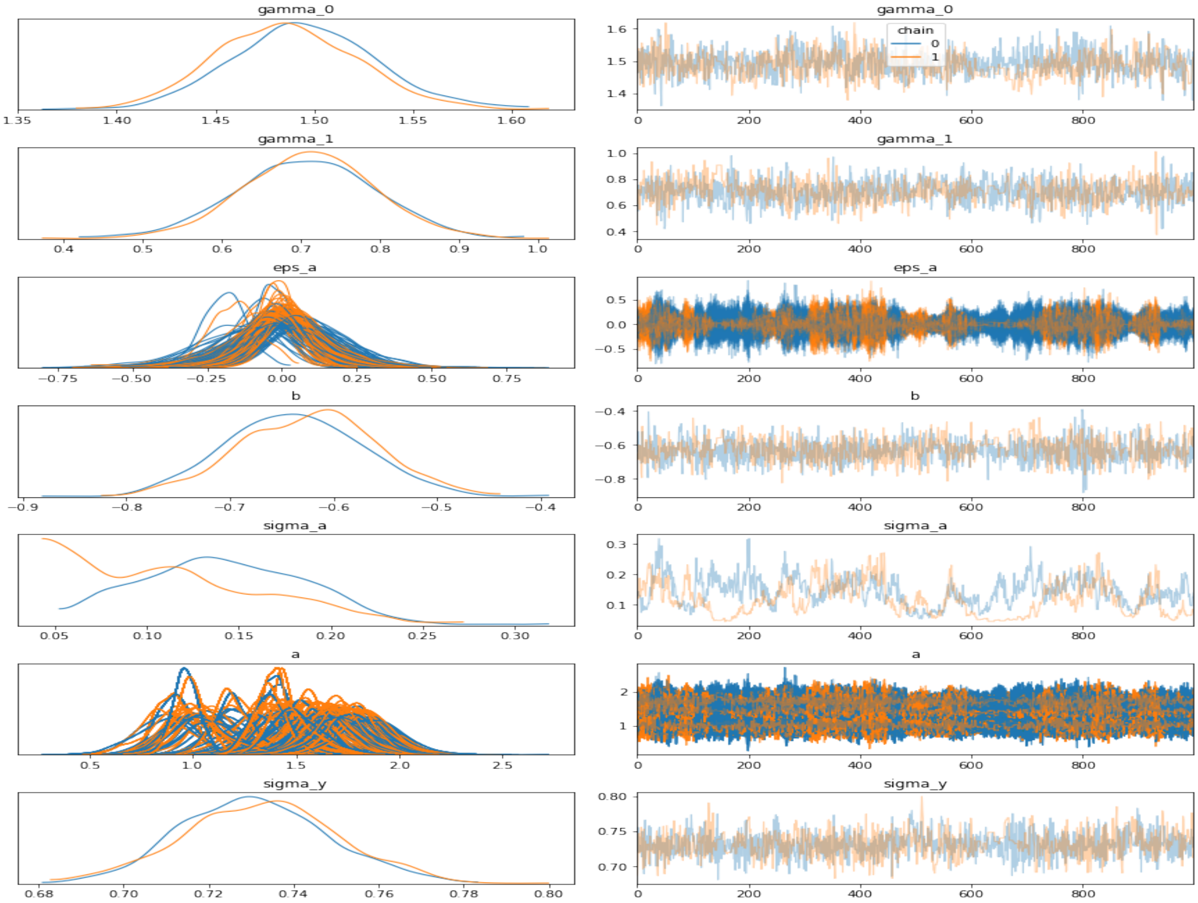
Two important predictors: measurement in basement or first floor (radon higher in basements), county uranium level (positive correlation with radon levels). The first step is to get county uranium by adding another datasets. Then we take log of the radon to get response variable.
In raw data, we can observe that County 'Lake' has the lowest radon count while 'Lac Qui Parle' has the highest average radon level.
For the modeling part, i am using the varying-intercept model as described in the post, the model architecture is listed below:



The predicted lowest county is 'Koochiching' and highest county is 'Rock', which does not match with actual perfectly.



The trace plot shows that the sampled has converged, as different chains demonstrated similar results and variations is relatively small.



10 Interview Questions

(a) Explain SVMs in your own words. Make sure you address:
(a) How are kernels introduced in SVMs
(b) How would you control for the slack variable penalty
(c) Why are slack variables not a problem when using a radial basis kernel
(d) What is the primal-dual relationship for finding the support vectors
(e) What is the run-time for optimization when using quadratic solvers, what happens as n

Ans:

(a) The function of kernel is to take data as input and transform it into the required form.The kernel function is what is applied on each data instance to map the original non-linear observations into a higher-dimensional space in which they become separable. Different SVM algorithms use different types of kernel functions. For example linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid. The kernel functions return the inner product between two points in a suitable feature space. Thus by defining a notion of similarity, with little computational cost even in very high-dimensional spaces.

(b) Slack variables are positive (or zero), local quantities that relax the stiff condition of linear separability, where each training point is seeing the same marginal hyper plane. slack variables can be geometrically defined as the ratio between the distance from a training point to a marginal hyperplane, and half of the margin. Large penalty will reduce the margin while small penalty tends to increase the margin.

(c) Since the value of the RBF kernel decreases with distance and ranges between zero (in the limit) and one, it has a ready interpretation as a similarity measure. It is a stationary kernel, which means that it is invariant to translation. The kernel function can be thought of as a cheap way of computing an infinite dimensional inner product

(d) For SVM the objective is to $min||w^2|| + c\sum_{i=1} n\zeta_i$ where ζ is the slack variable.we can re-wrie the form as $||w^2|| = \sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i^T x_j) + b*$ In such a case, the dual form of SVMs will be:

$$w_{\alpha \geq 0} \sum_i \alpha_i - 1/2 \sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

This optimization problem can be solved using quadratic programming, which is simple enough. The α terms can be interpreted as support vectors , which in practice is usually a sparse subset of the data. Finally, since we can express the optimization problem and classification function in terms of dot products with the training data, SVMs lend themselves naturally to kernels, which allow us to perform linear classification in high (potentially infinite) dimensional spaces

(e) It grows at least like n^2 when C is small and n^3 when C gets large

(b) Show that for the class of distributions in the regular exponential family that the mean update function is a weighted average of the prior distribution and observations.
Ans:

(c) Why do hierarchical models provide better model fits and regularization when data is sparse
Ans:
An advantage to using hierarchical models is their flexibility in modeling the continuum from all groups have the same parameters to all groups have completely different parameters. If the means of each group are actually similar (or identical) then σ^2 will be estimated to be small and the resulting inference for the individual θ will be almost the same as if you had just assumed a common mean for all groups. In contast, if the groups have very different means, then σ^2 will be large and the resulting inference for the individual θ will be almost the same as you didn't have the hierarchical model at all. Thus you didn't have to choose whether to use a model with a common mean for all groups or a completely independent mean for all groups, the hierarchical model allowed the data to tell you where you fell along that continuum.

An additional advantage to hierarchical models occurs when the number of observations for groups varies widely. In these situations, the groups with smaller numbers of observations will have improved inference about their group parameters by borrowing information via the hierarchical model about the group specific parameters. It is also the reason why hierarchical models provide better model fits and regularization when data is sparse.