General vs. personalized modeling: When and how to get specific

Adina Lederhendler, Senior data scientist, Neura

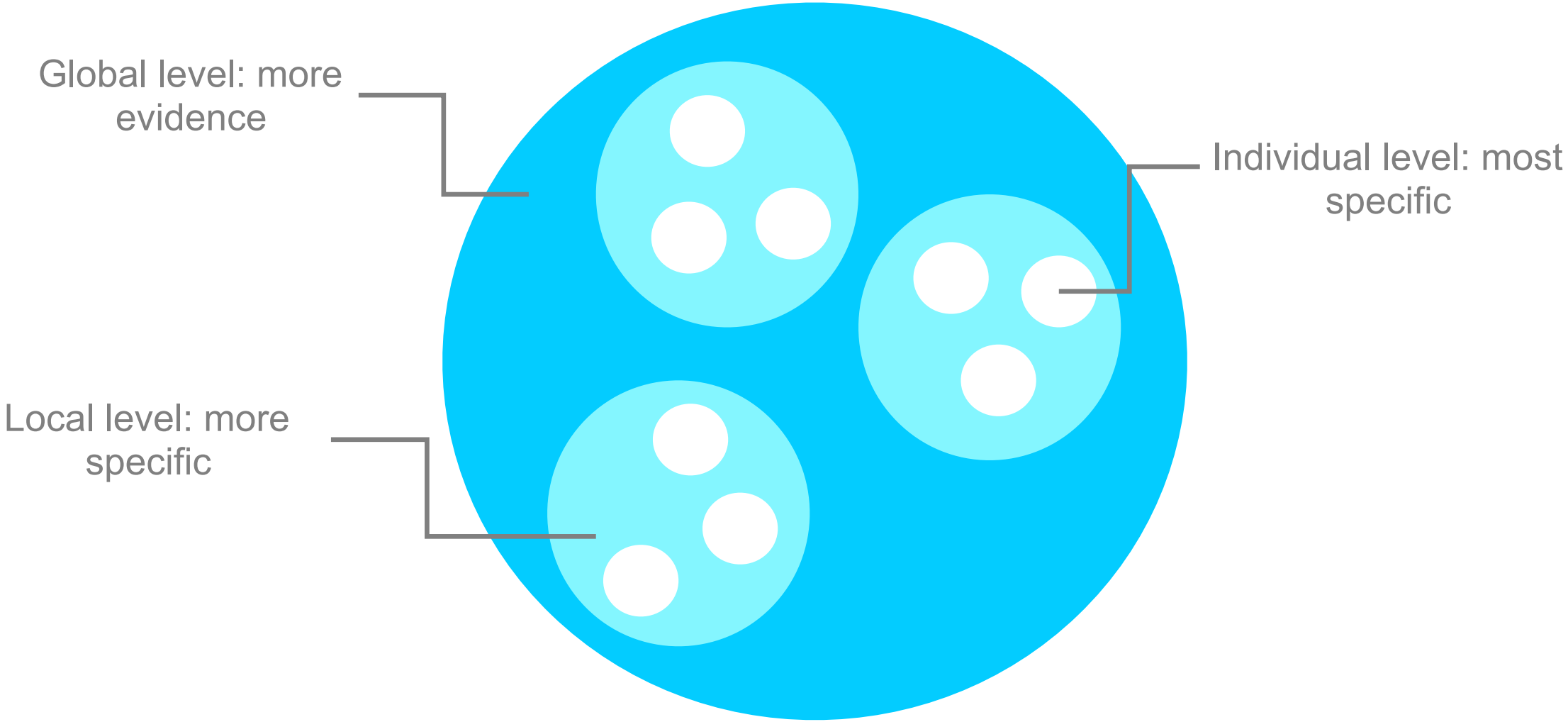
NEURA

An AI service enabling any app or device to boost user engagement by adapting to each user’s persona & lifestyle and reacting to key moments throughout their day.

Outline

- Global vs Personalized modeling different considerations
- Local Level Models subpopulation segmentation
- Hierarchical Models personalized models with shared parameters

Global vs Personalized modeling general considerations



NEURA

Global vs Personalized modeling data considerations

Global model - Risk of bias

Does the population represent the individuals well?

- Is it reasonable to assume all individuals share one distribution?
- Are certain subpopulations missing certain features?
- Do individuals' data change more rapidly than the population average?

Individual models - Risk of overfitting

Is there enough data to train a model at the individual level?

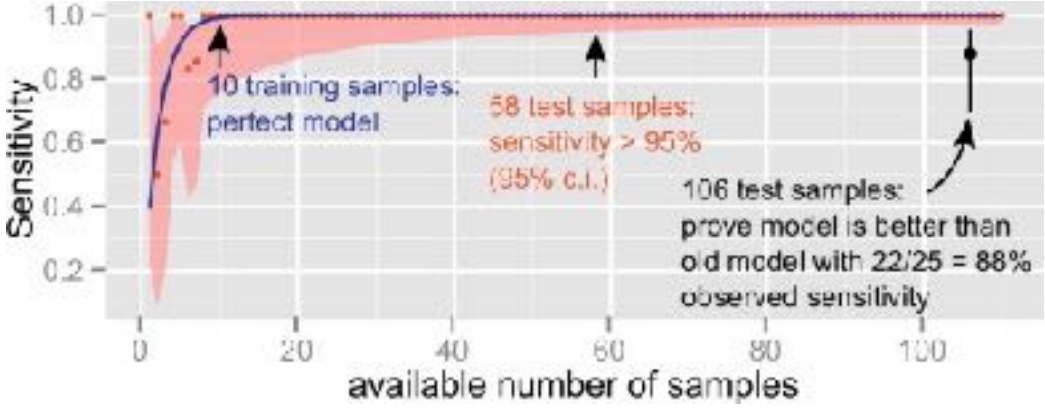
- How many data points per individual?
- How much noise in the data – can you trust individual measurements?

NEURA

Global vs Personalized modeling data considerations

In most cases the answer is not straightforward

- Many “rules of thumb” for different types of models – e.g. the number of data points should be at least 10 times the number of parameters for regression models.
- You can gain a lot of insight by doing some empirical analysis:
  - Look at the feature distribution over the population
  - For time series data – look at dynamics of population average vs individuals
  - Construct a learning curve:



Sample Size Planning for Classification Models, http://www.sciencedirect.com/science/article/pii/S0003267012016479?via%3Dihub (https://arxiv.org/abs/1211.1323)

NEURA

Global vs Personalized modeling domain considerations

What is the right population for your particular problem?

- Is the average relevant at the individual level?
  - Prediction based on behavioral patterns – e.g. most people grocery shop on weekends, but a particular individual may shop at a different time.
- How important is individual accuracy? **Is it ok to consider variation between individuals as "noise"?**
  - Medical studies intentionally look for average effects over population.
  - Insurance/credit risk models – Main concern is with overall loss, although to optimize overall loss, individual accuracy is important.
  - Personalized recommendation models – only have value if they have high accuracy at the individual level.
- Will the model have to be applied to new individuals?
- Does the model need to be evaluated/trained online? Does the global population scale support that?

NEURA

Local Level Models subpopulation segmentation

Individuals are best represented by a similar subset

Each subset is independent and trained separately.

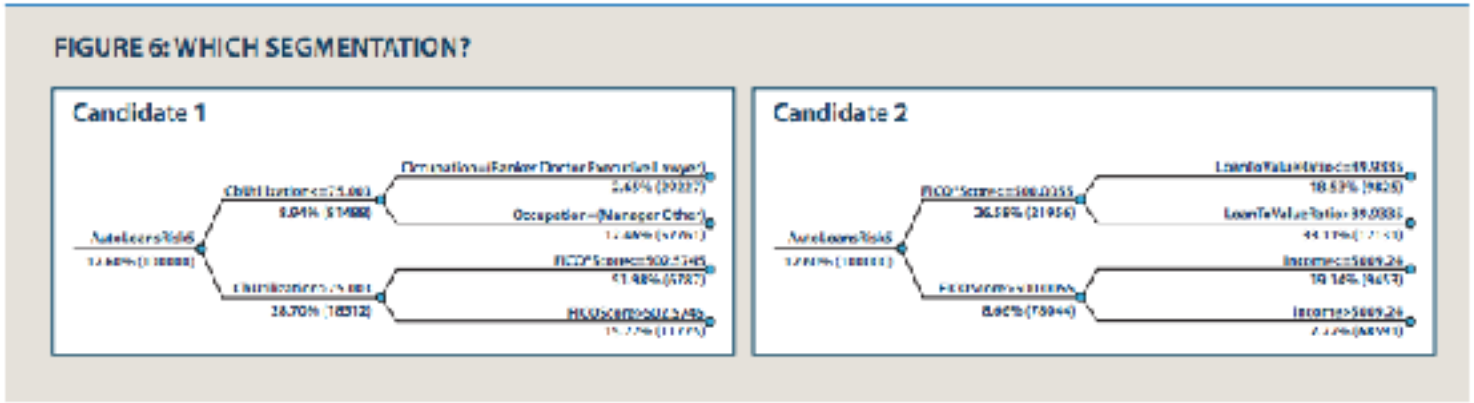
Most effective when:

- Natural segmentation or clustering
- Different features may be more relevant for different subsets

Local Level Models some examples

Segmented Scorecard module of FICO® Model Builder

- Automated module for choosing best population segmentation for credit risk scorecards.
- The system builds multiple decision trees:



- Chooses best tree based on performance of scorecards at each leaf.

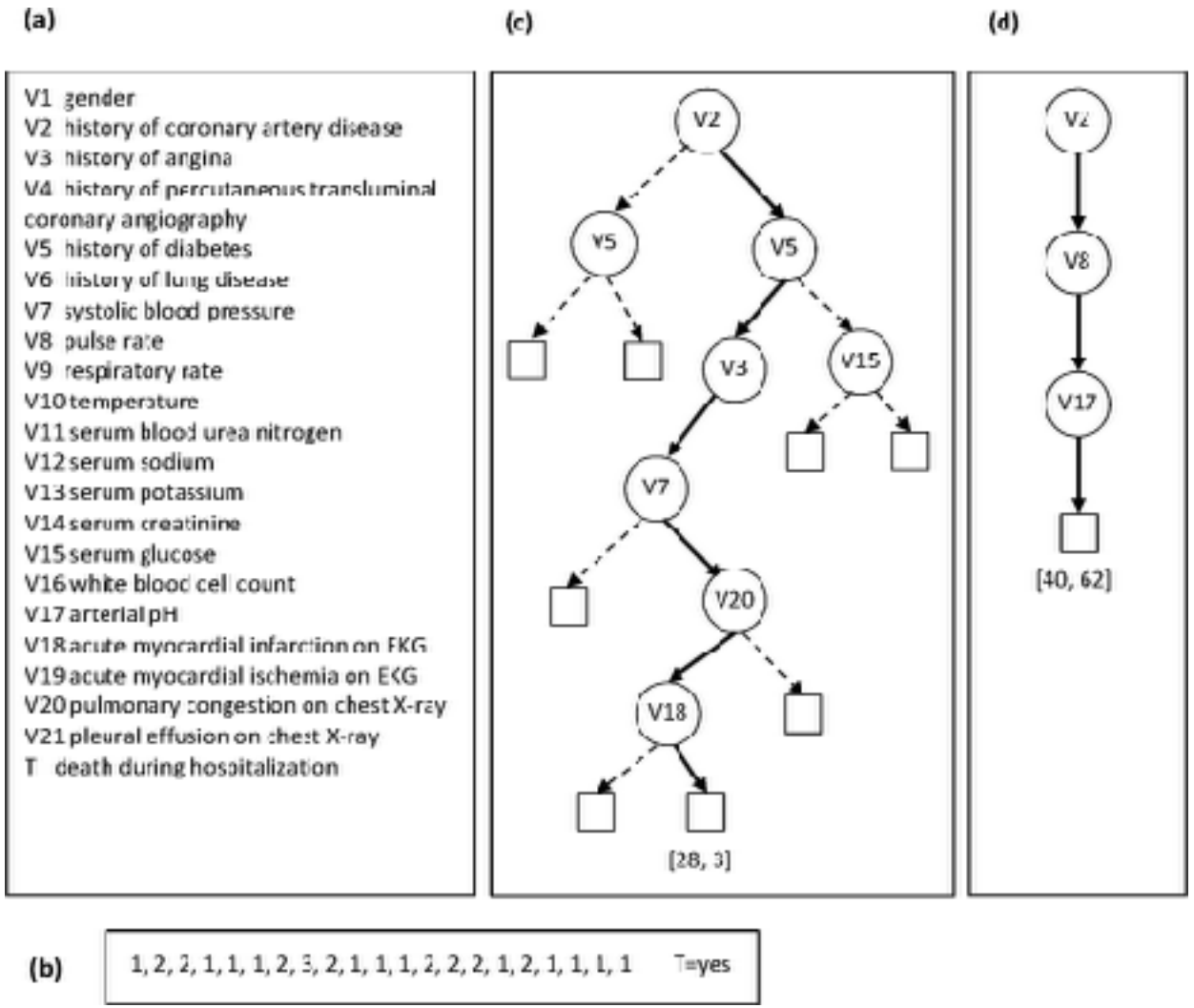
http://www.fico.com/en/node/8140?file=9737

NEURA

Local Level Models some examples

Decision-Path Model

- Instance-based model
- Binary outcome T
- The goal is to construct a “personalized” decision tree that is most suited to predict outcome T for a particular patient (LazyDT).
- At each node choose the feature that optimizes a score (e.g. information gain) out of the particular patient’s features.
- At the end, each path is pruned to the node with the highest score.



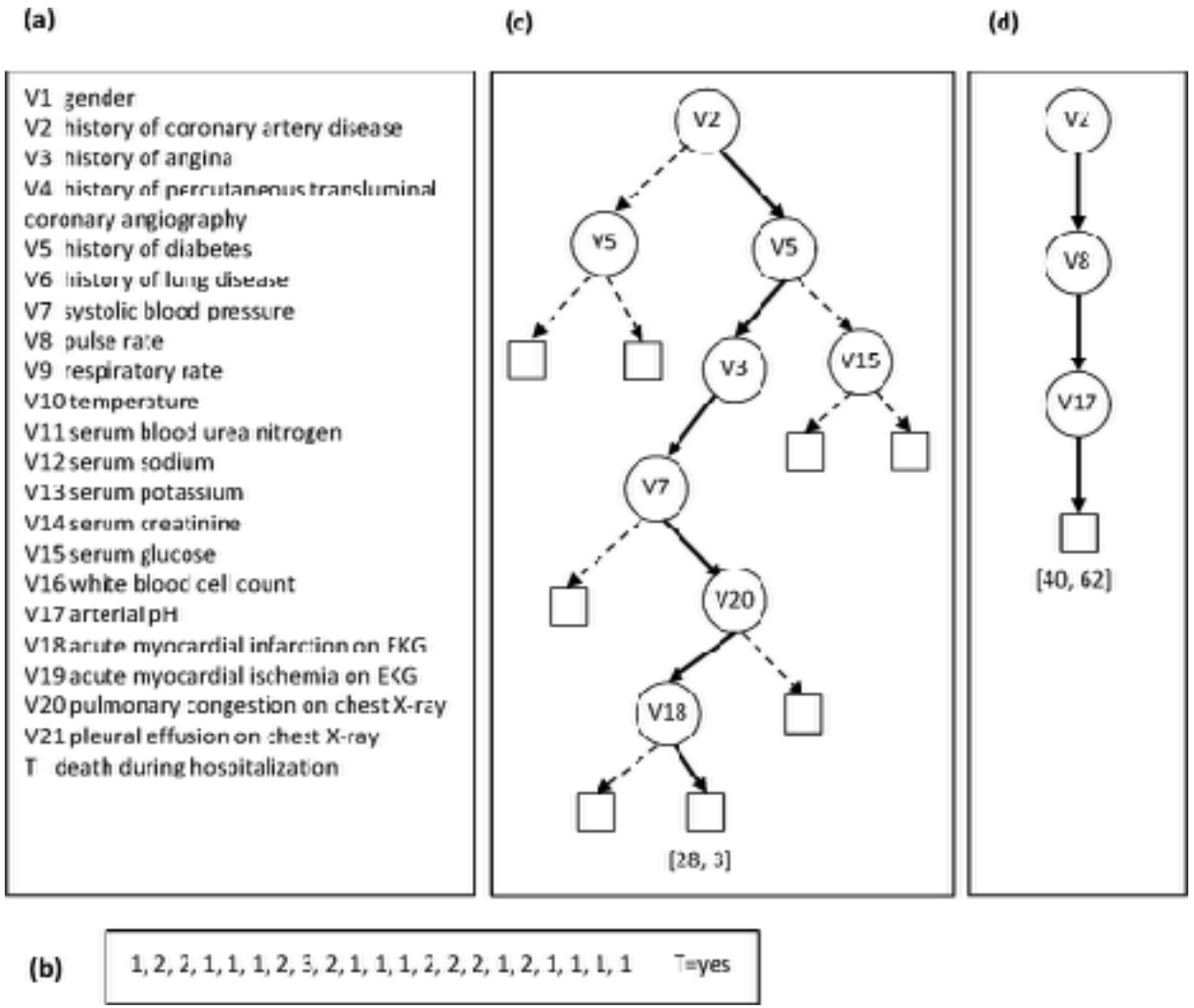
Personalized Modeling for Prediction with Decision-Path Models
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0131022

NEURA

Local Level Models some examples

Decision-Path Model

- Personalization comes from:
  - Only test patient's features are considered – good for sparse data sets.
  - If the score is not averaged over all sides of the split, it gives a chance to features that are rare but significant for a small population. (the authors of this paper chose to average over both sides of the split)
  - Pruning each path separately



Personalized Modeling for Prediction with Decision-Path Models
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0131022

NEURA

Hierarchical Models personalized models with shared parameters

Individuals are diverse but not independent

Use information from more than one level simultaneously, population and individual are modeled together.

Most effective when:

- Models reflect hierarchy in the population
- Some features are shared by many individuals

NEURA

Hierarchical Models hierarchical Bayes method

Hierarchical Bayes Method

- The distribution of individual parameters within a group is modeled by a higher level distribution with its own parameters (or many levels).
- Individual and group-level parameters are evaluated simultaneously. That way each individual's parameters are informed by the group.

Hierarchical Models hierarchical Bayes method

Classic example

Heads of a coin depends on bias ($p(heads) = \theta$), which in turn depends on coin manufacturing process. Say we have J coins from the same manufacturing line, each tossed n_j times resulting in k_j heads.

$$P(k_j|\theta_j, n_j) = \binom{n_j}{k_j} \theta_j^{k_j} (1-\theta_j)^{n_j - k_j} \qquad j = 1, \dots, J$$

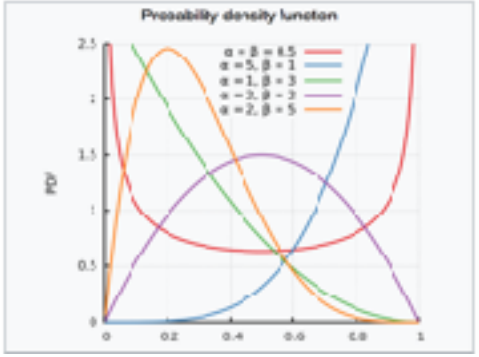
Model assumption: Each coin's bias θ_j is drawn from a prior distribution over all coins manufactured on the same line.

NEURA

Hierarchical Models hierarchical Bayes method

Choose a Beta distribution prior for θ_j, where α and β are shared by all coins in the manufacturing line.

$$\pi(\theta|\,\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\underbrace{\int_0^1 d\theta\,\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{B(\alpha,\beta)}} = Beta(\alpha,\beta)$$



mean

$$\mu = \frac{\alpha}{\alpha+\beta}$$

α and β can be estimated by maximizing the likelihood for the overall sample:

$$L(\boldsymbol{k}|\alpha,\beta,\boldsymbol{n}) = \prod_j \int_0^1 d\theta\, P(k_j|\theta,n_j)\pi(\theta|\,\alpha,\beta)$$

$$= \prod_j \binom{n_j}{k_j} \frac{B(k_j+\alpha, n_j-k_j+\beta)}{B(\alpha,\beta)}$$

NEURA

Hierarchical Models hierarchical Bayes method

Given α and β we can estimate the posterior distribution for θ_j:

$$g(\theta_j|\,\alpha,\beta,k_j,n_j) = \frac{P(k_j|\theta,n_j)\pi(\theta|\,\alpha,\beta)}{\int_0^1 d\theta\, P(k_j|\theta,n_j)\pi(\theta|\,\alpha,\beta)} = Beta(k_j+\alpha, n_j-k_j+\beta)$$

And so for coin *j*:

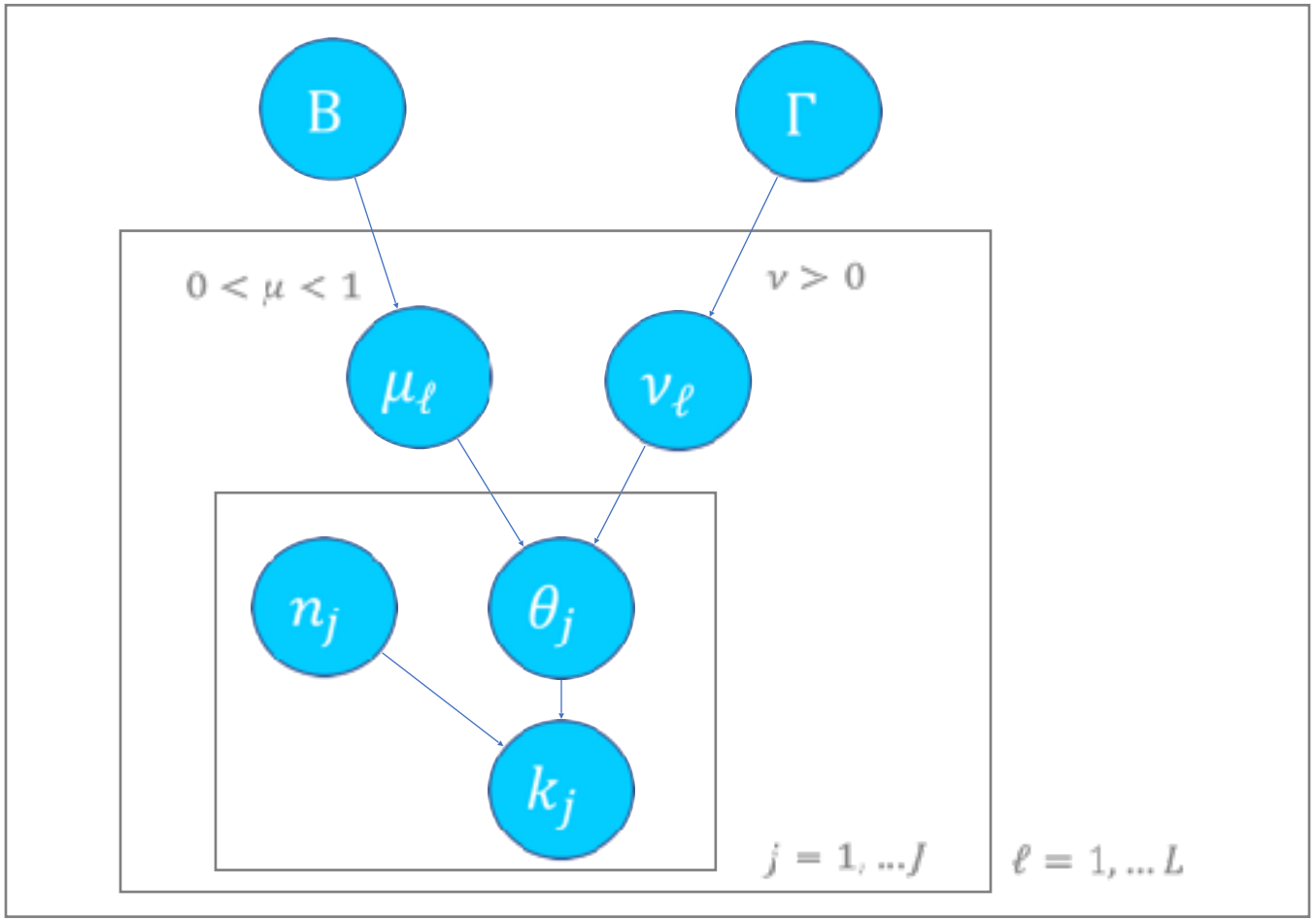
$$E_g[\theta_j] = \frac{k_j+\alpha}{n_j+\alpha+\beta} = \underbrace{\left(\frac{k_j}{n_j}\right)}_{\text{Sample average for coin } j}\underbrace{\left(\frac{n_j}{n_j+\alpha+\beta}\right)}_{\text{“Weight” of sample}} + \underbrace{\left(\frac{\alpha}{\alpha+\beta}\right)}_{\text{Prior mean}}\underbrace{\left(\frac{\alpha+\beta}{n_j+\alpha+\beta}\right)}_{\text{“Weight” of prior}}$$

$\nu = \alpha + \beta$ represents an effective sample size for the prior distribution

NEURA

Hierarchical Models hierarchical Bayes method

If there is more than one manufacturing line, we add another layer to the hierarchy:



All parameters are evaluated simultaneously from samples of all coins from all lines.

NEURA

Hierarchical Models multilevel features

Classic example: home radon levels

Model of home radon levels based on one individual home level feature, x_{ij} (whether the measurement of house *i* in county *j* was taken in the basement), and one county level feature, u_j (county-level uranium level).

$$y_{ij} \sim N(\alpha_j + \beta x_{ij}, \sigma_y^2) \quad i = 1, \dots, n_j, j = 1, \dots, J$$
$$\alpha_j \sim N(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2) \quad j = 1, \dots, J$$

- σ_y^2 - variation within county not explained by basement indicator, shared by all individuals
- σ_α^2 - variation between counties not explained by uranium, shared by all counties
- α_j, β - also shared

Multilevel (Hierarchical) Modeling: What It Can and Cannot Do, http://www.stat.columbia.edu/~gelman/research/published/multi2.pdf

NEURA

Hierarchical Models multitask regression

Age estimation

- Problem of trying to estimate person's age from photograph. Model based on two available databases of tagged photos, several per user.
- Previous studies show that personalized models perform better than global models.
- The authors propose a multi-level expansion of a Warped Gaussian Process

Figure 4. Sample images of one person in the FG-NET database.

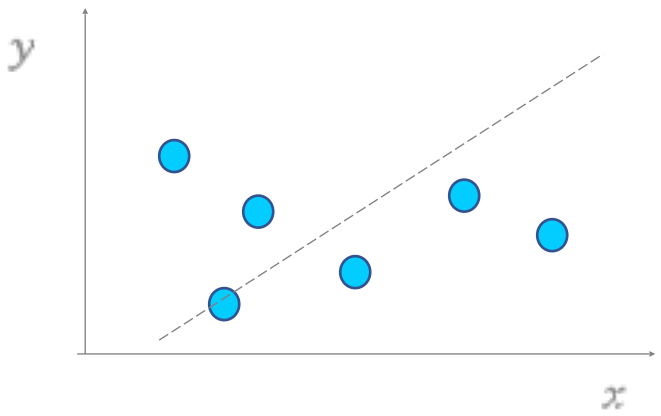
Figure 5. Sample images of two persons in the MORPH database.

Multi-Task Warped Gaussian Process for Personalized Age Estimation, http://repository.ust.hk/ir/bitstream/1783.1-6695/1/multi_taskw.pdf

NEURA

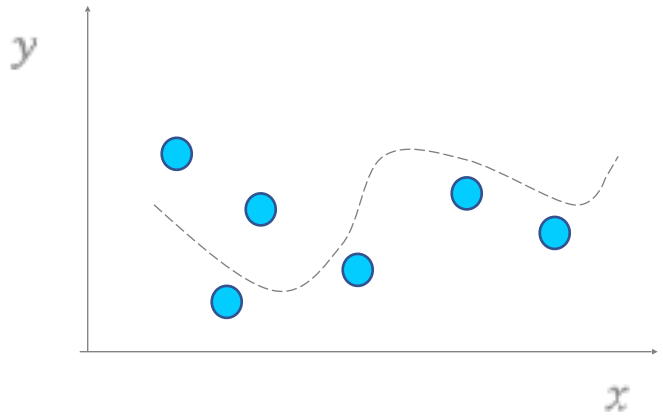
Hierarchical Models multitask regression

Gaussian Process (GP)



$$y = \theta_0 + \theta_1 x + \epsilon$$

Bayesian method estimates the distribution of parameters that best fits the observed data.



$$y = f(x) + \epsilon$$

Gaussian process estimates the distribution of all possible functions $f(x)$.

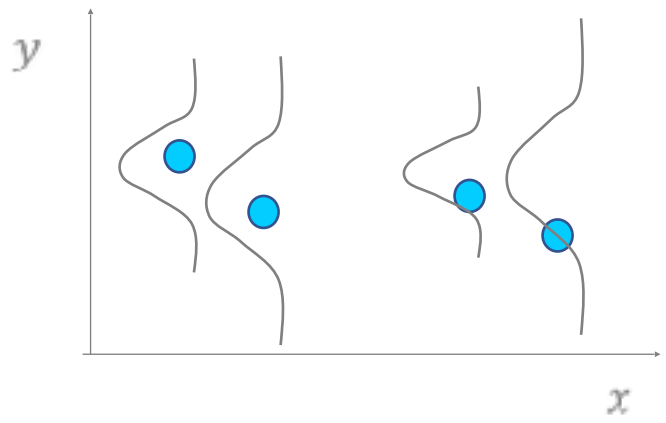
Multi-Task Warped Gaussian Process for Personalized Age Estimation, http://repository.ust.hk/ir/bitstream/1783.1-6695/1/multi_taskw.pdf

NEURA

Hierarchical Models multitask regression

Gaussian Process (GP)

GP represents the function $f(x)$ as a series of values f_i which are jointly Gaussian with correlated errors



f is modeled as a series of Gaussian variables at each x

$$p(\boldsymbol{f}|\boldsymbol{x}) \sim N(\mathbf{0}_n, \mathbf{K})$$

$$p(y_i|f_i) \sim N(f_i, \sigma^2)$$

Parameters θ, σ are estimated from the likelihood

$$p(\boldsymbol{y}|\boldsymbol{x}) \sim N(\mathbf{0}_n, \mathbf{K} + \sigma^2 \mathbf{I}_n)$$

And then f can be evaluated on new x_*

fs are correlated through the covariance matrix $\mathbf{K}$:

$$K_{ij} = k_\theta(x_i, x_j)$$

k_θ is the kernel function defining similarity between xs.

-> If x's are similar then corresponding f's are also similar.

$\mathbf{K}$ and σ^2 contain all the information of how y depends on x and are learned automatically from the data

Multi-Task Warped Gaussian Process for Personalized Age Estimation, http://repository.ust.hk/ir/bitstream/1783.1-6695/1/multi_taskw.pdf

NEURA

Hierarchical Models multitask regression

Warped Gaussian Process (WGP)

A generalization of a GP where the outputs $\{y_i\}$ do not satisfy the GP assumptions, but their transformations $z_i = g_\phi(y_i)$ do.

$$p(f|\boldsymbol{x}) \sim N(\mathbf{0}_n, \mathbf{K})$$

$$p(z_i = g_\phi(y_i)|f_i) \sim N(f_i, \sigma^2)$$

Multi-Task Warped Gaussian Process for Personalized Age Estimation, http://repository.ust.hk/ir/bitstream/1783.1-6695/1/multi_taskw.pdf

NEURA

Hierarchical Models multitask regression

Multi-Task WGP

- Multi-task learning is a method of learning several models ("tasks") at once. Sharing information (e.g. structure) between tasks helps improve their accuracy.
- Hierarchical models can be treated as multi-task models where each individual is a "task".
- Each individual i has n_i data points (photos) $\{(\boldsymbol{x}_j^i, y_j^i)\}_{j=1}^{n_i}$

  $\boldsymbol{x}_j^i$ - feature vector from photo j of individual i

  y_j^i - age of individual i in photo j

Multi-Task Warped Gaussian Process for Personalized Age Estimation, http://repository.ust.hk/ir/bitstream/1783.1-6695/1/multi_taskw.pdf

NEURA

Hierarchical Models multitask regression

Multi-Task WGP

- The data is modeled as a WGP

$$p(\boldsymbol{f}|\boldsymbol{X}) \sim N(\mathbf{0}_n, \mathbf{K})$$

[illegible]

- The likelihood of each data point is defined on a latent variable $z_j^i = g_\phi(y_j^i)$

$$p(z_j^i|y_j^i) \sim N(y_j^i, \sigma_i^2)$$

- $\mathbf{K}$, parameterized by θ and ϕ are shared by all tasks and capture the dependence of y on global features.
- σ_i^2 captures the dependence of y on "task" (individual) level features.

Multi-Task Warped Gaussian Process for Personalized Age Estimation, http://repository.ust.hk/ir/bitstream/1783.1-6695/1/multi_taskw.pdf

NEURA

Questions? Thank you

NEURA