Appendix B

Bayesian Gaussian Process Regression

The aim of a Bayesian Gaussian process regression is to estimate the likelihood that a given occupation can be automated, based on its skill content. The O*Net database contains detailed skills information for 954 occupations, including the level of each skill required and the importance of that skill to the job. The conceptual basis for the model is that there are some skills where humans still dominate robots, and that jobs that require high levels of these skills therefore cannot be fully automated. Conversely, jobs that do not require these skills, or where the skill is not important to the job, are more likely to be automatable.

The skill variables described in Table 1 fully specify an explanatory variable set or a *feature vector* for each occupation. The dependent variable (whether or not a job is automatable) is incomplete and, for the most part, unknown. We only know which jobs are fully automatable and, with some certainty, the jobs that are currently impossible to automate. For those jobs that are partially automatable, the level is unknown. To handle this difficulty with the data, we employed a Bayesian Gaussian process regression, a powerful kernel-based classification method that uses "training data" – the information about which we are certain – to determine the remaining unknown values probabilistically. This method also does not limit the relationship between skills and the potential for automation to one that is constant – that is, how much a certain skill affects the chances of being automated is not constant over all automation levels.

A Gaussian process is defined as a collection of random variables, any finite number of which have a joint Gaussian distribution (see Rasmussen and Williams 2006, 14). Gaussian processes have been extensively used for many different variants of non-parametric estimation (Seeger 2004).

To develop a training data set, we labelled occupations as either fully automatable ($y_i = 1$) or not automatable ($y_i = 0$) for a subset of 109 occupations ($\{y_i = 1{:}n = 65\}$, $\{y_i = 0{:}n = 44\}$), based on existing information and consultation with subject matter experts. The training data set is defined by $\mathcal{D} = (X,y)$, where X is the matrix of feature vectors and y gives the associated class label. Each element $x_{i,j}$ of X represents the level of each skill (j) required, weighted by the importance of that skill to the specified occupation (i). We assume that $\mathcal{D} = (X,y)$, $x_i \in \mathrm{R}^9$, $y_i \in \{0,1\}$ is a noisy independent, identically distributed sample from latent function $f{:}x \mapsto \mathrm{R}$, where $w = P(y|f)$ denotes the noise distribution.

Given dataset $\mathcal{D}$ of n observations, $\mathcal{D} = \{(x_i, y_i)|\mathrm{i} = 1,2,\ldots,n\}$, we wished to make predictions (y_*) for new input features X_* that are not in the training set $\mathcal{D}$. To estimate y_*, we used a zero-mean Gaussian process prior ($w \sim N(0, \Sigma)$) and a generative Bayesian method. The process gives a prior probability weight to every possible function that describes the relationship specified by $\mathcal{D}$, where higher probabilities are assigned to functions we consider more likely. The likelihood of a function is determined by its relative proximity to training data points and the prior specification that fixes the properties of functions to be considered for inference.

The model is then computed by:

1 Introducing $\phi(x)$, which maps input vector x into an N-dimensional feature space. Further, let $\Phi(X)$ be the aggregation of the columns $\phi(x)$. The model is defined by:

$$f(x) = \phi(x)^T w.$$

2 Conditioning the prior distribution on the data $\mathcal{D}$, resulting in a posterior distribution:

$$posterior = \frac{likelihood\ x\ prior}{marginal\ likelihood}, \qquad p(w|y,\Phi(X)) = \frac{p(y|\Phi(X),w)p(w)}{p(y|\Phi(X))}$$

$$p(w|y,\Phi(X)) \sim N(\bar{w} = \frac{1}{\sigma_n^2} A^{-1}\Phi(X)y,\ A^{-1}),$$

where $A = \sigma_n^{-2}\,\Phi(X)\Phi(X)^T + \Sigma^{-1}$.

3 The marginal posterior is the predictive distribution

$$f_*|x_*,\mathcal{D} = \int p(f_*|\phi_*,w)p(w|\Phi,y)dw$$
$$\sim N(\frac{1}{\sigma_n^2}\,\phi_*^T A^{-1}\Phi y, \phi_*^T A^{-1}\phi_*)$$

where $\Phi = \Phi(X)$ and $\phi_* = \phi(x_*)$.

An alternative formulation is given by

$$f_*|x_*,\mathcal{D} \sim N(\phi_*^T\Sigma\Phi(K + \sigma_n^2 I)^{-1}y,\ \phi_*^T\Sigma\phi_* - \phi_*^T\Sigma\Phi(K + \sigma_n^2 I)^{-1}\Phi^T\Sigma\phi_*),$$

where $K = \Phi^T\Sigma\Phi$.

Notice that, in the equation above, the feature space always enters in a form that dictates entries in the matrices are of the form $\phi(x)^T\Sigma\phi(x')$, where x and x' are in either the training set or the test set.

Let $k(x,x') = \phi(x)^T\,\Sigma\phi(x')$ be the kernel or covariance function. The specification of a covariance function implies a distribution over functions. A Gaussian process is completely specified by its mean function and covariance function:

$$f(x) \sim GP(m(x),k(x,x'))$$
$$y = f(x) + \varepsilon.$$

For our model $f(x)=\phi(x)^Tw$ with prior $w\sim N(0,\Sigma)$ we have mean and covariance

$$E[f(x)] = 0$$
$$E[f(x)f(x')] = \phi(x)^T\,\Sigma\phi(x') = k(x,x').$$

Notice that the covariance between outputs is a function of inputs. We defined the kernel as the squared exponential covariance function:

$$k(x,x') = \exp\,(-\frac{1}{2}|x - x'|^2).$$

Assuming additive, identically distributed Gaussian noise ε with variance σ_n^2, the prior on noisy observations becomes:

$$cov(y) = K(X,X) + \sigma_n^2 .$$

The free parameter σ_n^2 is referred to as a hyperparameter to emphasize that it refers to a parameter of a non-parametric model. The parameters (weights) of the underlying parametric model have been integrated out.

To compute the model, we used the "kernlab" R package (Karatzoglou, Smola and Hornik 2016). Specifically, we employed the "gausspr" function, an exponentiated quadratic covariance (radial basis function kernel = "rbfdot") and an optimized hyperparameter selection ($\sigma_n^2 = 0.13$). The accuracy of the model is estimated by 10-fold cross validation error, which yields an error estimate of 0.10 on predicted probabilities.

Threshold Values for Risk Categorization

To determine an appropriate threshold for high, medium and low risk of automation categories we take into account the differences in distribution, conditional on occupation type. Some job types are heavily skewed towards low risk of automation (eg. Management occupations) and some are heavily skewed towards high risk of automation (eg. Manufacturing occupations). To account for dispersion differences in the underlying distributions of each job type, the standard deviation, proportioned by the skew, is estimated for each distribution to calculate a range around the mean of each distribution. These ranges are then weighted by the number of observations in each occupation type to determine a single aggregate range of "medium risk" of automation.

Threshold measures for each distribution are calculated as follows:

$$\mu \pm \mu_2\mu_3$$

$$= \mu \pm (E[(X - \mu)^2]) \bullet \frac{E[(X - \mu)^3]}{E[(X - \mu)^2]^{\frac{3}{2}}}$$

$$= \mu \pm \frac{E[(X - \mu)^3]}{E[(X - \mu)^2]^{\frac{1}{2}}}$$

$$= \mu \pm \left(\frac{E[X^3] - 3\mu E[X^2] + 3\mu^2 E[X] - \mu^3}{E[X^2] - \mu^2}\right)$$

$$= \mu \pm \left(\frac{E[X^3] - 3\mu(E[X^2] - \mu E[X]) - \mu^3}{E[X^2] - \mu^2}\right)$$

$$= \mu \pm \left(\frac{E[X^3] - \mu^3}{E[X^2] - \mu^2} - 3\mu\right)$$

$$= 4\mu - \frac{E[X^3] - \mu^3}{E[X^2] - \mu^2}, -2\mu + \frac{E[X^3] - \mu^3}{E[X^2] - \mu^2}$$

For each job type we now have:

$$(c_{j,low}, c_{j,high}) = \begin{cases} (4\mu - \frac{E[X^3] - \mu^3}{E[X^2] - \mu^2}, -2\mu + \frac{E[X^3] - \mu^3}{E[X^2] - \mu^2}) & \text{if } \mu_3 > 0 \\ (-2\mu + \frac{E[X^3] - \mu^3}{E[X^2] - \mu^2}, 4\mu - \frac{E[X^3] - \mu^3}{E[X^2] - \mu^2}) & \text{otherwise} \end{cases}$$

$$= \begin{cases} (\mu - \mu_2\mu_3, \mu + \mu_2\mu_3) & \text{if } \mu_3 > 0 \\ (\mu + \mu_2\mu_3, \mu - \mu_2\mu_3) & \text{otherwise} \end{cases}$$

We now have threshold measures for each underlying distribution:

$C = [c_{j,low}, c_{j,high}] \mid j \in \{A,B,C,D,E,F,G,H,I,J\}$.

The threshold measures for the aggregate distribution are then:

$$\begin{aligned} \textit{Risk threshold} &= [low, high] \\ &= \left[\frac{\Sigma_{j=A}^{J} c_{j,low} \cdot n_j}{\Sigma_{j=A}^{J} n_j}, \frac{\Sigma_{j=A}^{J} c_{j,high} \cdot n_j}{\Sigma_{j=A}^{J} n_j} \right] \\ &= [0.36, 0.72] \end{aligned}$$

This is equivalent to $E[\textit{Probability of Automation}] \pm 0.18$.