Joint Emotion Analysis via Multi-task Gaussian Processes

Daniel Beck[†] **Trevor Cohn**[‡] **Lucia Specia**[†]
[†]Department of Computer Science, University of Sheffield, United Kingdom
{debeck1,l.specia}@sheffield.ac.uk
[‡]Computing and Information Systems, University of Melbourne, Australia
t.cohn@unimelb.edu.au

Abstract

We propose a model for jointly predicting multiple emotions in natural language sentences. Our model is based on a low-rank coregionalisation approach, which combines a vector-valued Gaussian Process with a rich parameterisation scheme. We show that our approach is able to learn correlations and anti-correlations between emotions on a news headlines dataset. The proposed model outperforms both single-task baselines and other multi-task approaches.

1 Introduction

Multi-task learning (Caruana, 1997) has been widely used in Natural Language Processing. Most of these learning methods are aimed for Domain Adaptation (Daumé III, 2007; Finkel and Manning, 2009), where we hypothesize that we can learn from multiple domains by assuming similarities between them. A more recent use of multi-task learning is to model annotator bias and noise for datasets labelled by multiple annotators (Cohn and Specia, 2013).

The settings mentioned above have one aspect in common: they assume some degree of positive correlation between tasks. In Domain Adaptation, we assume that some "general", domain-independent knowledge exists in the data. For annotator noise modelling, we assume that a "ground truth" exists and that annotations are some noisy deviations from this truth. However, for some settings these assumptions do not necessarily hold and often tasks can be *anti-correlated*. For these cases, we need to employ multi-task methods that are able to learn these relations from data and correctly employ them when making predictions, avoiding negative knowledge transfer.

An example of a problem that shows this behaviour is Emotion Analysis, where the goal is to automatically detect emotions in a text (Strapparava and Mihalcea, 2008; Mihalcea and Strapparava, 2012). This problem is closely related to Opinion Mining (Pang and Lee, 2008), with similar applications, but it is usually done at a more fine-grained level and involves the prediction of a set of labels (one for each emotion) instead of a single label. While we expect some emotions to have some degree of correlation, this is usually not the case for all possible emotions. For instance, we expect *sadness* and *joy* to be anti-correlated.

We propose a multi-task setting for Emotion Analysis based on a vector-valued Gaussian Process (GP) approach known as *coregionalisation* (Álvarez et al., 2012). The idea is to combine a GP with a low-rank matrix which encodes task correlations. Our motivation to employ this model is three-fold:

- Datasets for this task are scarce and small so we hypothesize that a multi-task approach will results in better models by allowing a task to borrow statistical strength from other tasks;
- The annotation scheme is subjective and very fine-grained, and is therefore heavily prone to bias and noise, both which can be modelled easily using GPs;
- Finally, we also have the goal to learn a model that shows sound and interpretable correlations between emotions.

2 Multi-task Gaussian Process Regression

Gaussian Processes (GPs) (Rasmussen and Williams, 2006) are a Bayesian kernelised framework considered the state-of-the-art for regression. They have been recently used successfully for translation quality prediction (Cohn and Specia, 2013; Beck et al., 2013; Shah et al., 2013)

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1798–1803,
October 25-29, 2014, Doha, Qatar.

and modelling text periodicities (Preotiuc-Pietro and Cohn, 2013). In the following we give a brief description on how GPs are applied in a regression setting.

Given an input $\mathbf{x}$, the GP regression assumes that its output y is a noise corrupted version of a latent function evaluation, $y = f(\mathbf{x}) + \eta$, where $\eta \sim \mathcal{N}(0, \sigma_n^2)$ is the added white noise and the function f is drawn from a GP prior:

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{1}$$

where $\mu(\mathbf{x})$ is the *mean* function, which is usually the 0 constant, and $k(\mathbf{x}, \mathbf{x}')$ is the kernel or *covariance* function, which describes the covariance between values of f at locations $\mathbf{x}$ and $\mathbf{x}'$.

To predict the value for an unseen input $\mathbf{x}_*$, we compute the Bayesian posterior, which can be calculated analytically, resulting in a Gaussian distribution over the output y_*:[1]

$$\begin{aligned} y_* \sim \mathcal{N}(&\mathbf{k}_*(\mathbf{K} + \sigma_n \mathbf{I})^{-1}\mathbf{y}^T, \\ &k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T(\mathbf{K} + \sigma_n \mathbf{I})^{-1}\mathbf{k}_*), \end{aligned} \tag{2}$$

where $\mathbf{K}$ is the Gram matrix corresponding to the covariance kernel evaluated at every pair of training inputs and $\mathbf{k}_* = [\langle \mathbf{x}_1, \mathbf{x}_* \rangle, \langle \mathbf{x}_2, \mathbf{x}_* \rangle, \ldots, \langle \mathbf{x}_n, \mathbf{x}_* \rangle]$ is the vector of kernel evaluations between the test input and each training input.

2.1 The Intrinsic Coregionalisation Model

By extending the GP regression framework to vector-valued outputs we obtain the so-called coregionalisation models. Specifically, we employ a separable vector-valued kernel known as *Intrinsic Coregionalisation Model* (ICM) (Álvarez et al., 2012). Considering a set of D tasks, we define the corresponding vector-valued kernel as:

$$k((\mathbf{x}, d), (\mathbf{x}', d')) = k_{\text{data}}(\mathbf{x}, \mathbf{x}') \times \mathbf{B}_{d,d'}, \tag{3}$$

where k_{data} is a kernel on the input points (here a Radial Basis Function, RBF), d and d' are task or metadata information for each input and $\mathbf{B} \in \mathbb{R}^{D \times D}$ is the coregionalisation matrix, which encodes task covariances and is symmetric and positive semi-definite.

A key advantage of GP-based modelling is its ability to learn hyperparameters directly from data by maximising the marginal likelihood:

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \int_f p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, f)p(f). \tag{4}$$

This process is usually performed to learn the noise variance and kernel hyperparameters, including the coregionalisation matrix. In order to do this, we need to consider how $\mathbf{B}$ is parameterised.

Cohn and Specia (2013) treat the diagonal values of $\mathbf{B}$ as hyperparameters, and as a consequence are able to leverage the inter-task transfer between each independent task and the global "pooled" task. They however fix non-diagonal values to 1, which in practice is equivalent to assuming equal correlation across tasks. This can be limiting, in that this formulation cannot model anti-correlations between tasks.

In this work we lift this restriction by adopting a different parameterisation of $\mathbf{B}$ that allows the learning of all task correlations. A straightforward way to do that would be to consider every correlation as an hyperparameter, but this can result in a matrix which is not positive semi-definite (and therefore, not a valid covariance matrix). To ensure this property, we follow the method proposed by Bonilla et al. (2008), which decomposes $\mathbf{B}$ using Probabilistic Principal Component Analysis:

$$\mathbf{B} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T + \text{diag}(\boldsymbol{\alpha}), \tag{5}$$

where $\mathbf{U}$ is an $D \times R$ matrix containing the R principal eigenvectors and Λ is a $R \times R$ diagonal matrix containing the corresponding eigenvalues. The choice of R defines the *rank* of $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, which can be understood as the capacity of the manifold with which we model the D tasks. The vector $\boldsymbol{\alpha}$ allows for each task to behave more or less independently with respect to the global task. The final rank of $\mathbf{B}$ depends on both terms in Equation 5.

For numerical stability, we use the incomplete-Cholesky decomposition over the matrix $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, resulting in the following parameterisation for $\mathbf{B}$:

$$\mathbf{B} = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T + \text{diag}(\boldsymbol{\alpha}), \tag{6}$$

where $\tilde{\mathbf{L}}$ is a $D \times R$ matrix. In this setting, we treat all elements of $\tilde{\mathbf{L}}$ as hyperparameters. Setting a larger rank allows more flexibility in modelling task correlations. However, a higher number of hyperparameters may lead to overfitting problems or otherwise cause issues in optimisation due

[1] We refer the reader to Rasmussen and Williams (2006, Chap. 2) for an in-depth explanation of GP regression.

to additional non-convexities in the log likelihood objective. In our experiments we evaluate this behaviour empirically by testing a range of ranks for each setting.

The low-rank model can subsume the ones proposed by Cohn and Specia (2013) by fixing and tying some of the hyperparameters:

Independent: fixing $\tilde{\mathbf{L}} = \mathbf{0}$ and $\boldsymbol{\alpha} = \mathbf{1}$;

Pooled: fixing $\tilde{\mathbf{L}} = \mathbf{1}$ and $\boldsymbol{\alpha} = \mathbf{0}$;

Combined: fixing $\tilde{\mathbf{L}} = \mathbf{1}$ and tying all components of $\boldsymbol{\alpha}$;

Combined+: fixing $\tilde{\mathbf{L}} = \mathbf{1}$.

These formulations allow us to easily replicate their modelling approach, which we evaluate as competitive baselines in our experiments.

3 Experimental Setup

To address the feasibility of our approach, we propose a set of experiments with three goals in mind:

- To find our whether the ICM is able to learn sensible emotion correlations;
- To check if these correlations are able to improve predictions for unseen texts;
- To investigate the behaviour of the ICM model as we increase the training set size.

Dataset We use the dataset provided by the "Affective Text" shared task in SemEval-2007 (Strapparava and Mihalcea, 2007), which is composed of 1000 news headlines annotated in terms of six emotions: *Anger, Disgust, Fear, Joy, Sadness* and *Surprise*. For each emotion, a score between 0 and 100 is given, 0 meaning total lack of emotion and 100 maximum emotional load. We use 100 sentences for training and the remaining 900 for testing.

Model For all experiments, we use a Radial Basis Function (RBF) data kernel over a bag-of-words feature representation. Words were downcased and lemmatized using the WordNet lemmatizer in the NLTK[2] toolkit (Bird et al., 2009). We then use the GPy toolkit[3] to combine this kernel with a coregionalisation model over the six emotions, comparing a number of low-rank approximations.

Baselines and Evaluation We compare prediction results with a set of single-task baselines: a Support Vector Machine (SVM) using an RBF kernel with hyperparameters optimised via cross-validation and a single-task GP, optimised via likelihood maximisation. The SVM models were trained using the Scikit-learn toolkit[4] (Pedregosa et al., 2011). We also compare our results against the ones obtained by employing the "Combined" and "Combined+" models proposed by Cohn and Specia (2013). Following previous work in this area, we use Pearson's correlation coefficient as evaluation metric.

4 Results and Discussion

4.1 Learned Task Correlations

Figure 1 shows the learned coregionalisation matrix setting the initial rank as 1, reordering the emotions to emphasize the learned structure. We can see that the matrix follows a block structure, clustering some of the emotions. This picture shows two interesting behaviours:

- *Sadness* and *fear* are highly correlated. *Anger* and *disgust* also correlate with them, although to a lesser extent, and could be considered as belonging to the same cluster. We can also see correlation between *surprise* and *joy*. These are intuitively sound clusters based on the polarity of these emotions.
- In addition to correlations, the model learns anti-correlations, especially between *joy/surprise* and the other emotions. We also note that *joy* has the highest diagonal value, meaning that it gives preference to independent modelling (instead of pooling over the remaining tasks).

Inspecting the eigenvalues of the learned matrix allows us to empirically determine its resulting rank. In this case we find that the model has learned a matrix of rank 3, which indicates that our initial assumption of a rank 1 coregionalisation matrix may be too small in terms of modelling capacity[5]. This suggests that a higher rank is justified, although care must be taken due to the local optima and overfitting issues cited in §2.1.

[2] http://www.nltk.org

[3] http://github.com/SheffieldML/GPy

[4] http://scikit-learn.org

[5] The eigenvalues were $592, 62, 86, 4, 3 \times 10^{-3}$ and 9×10^{-5}.

	Anger	Disgust	Fear	Joy	Sadness	Surprise	All
SVM	**0.3084**	**0.2135**	0.3525	0.0905	**0.3330**	0.1148	0.2603
Single GP	0.1683	0.0035	0.3462	0.2035	0.3011	0.1599	0.3659
ICM GP (Combined)	0.2301	0.1230	0.2913	0.2202	0.2303	0.1744	0.3295
ICM GP (Combined+)	0.1539	0.1240	0.3438	0.2466	0.2850	0.2027	0.3723
ICM GP (Rank 1)	0.2133	0.1075	0.3623	**0.2810**	0.3137	0.2415	**0.3988**
ICM GP (Rank 5)	0.2542	0.1799	**0.3727**	0.2711	0.3157	**0.2446**	0.3957

Table 1: Prediction results in terms of Pearson's correlation coefficient (higher is better). Boldface values show the best performing model for each emotion. The scores for the "All" column were calculated over the predictions for all emotions concatenated (instead of just averaging over the scores for each emotion).

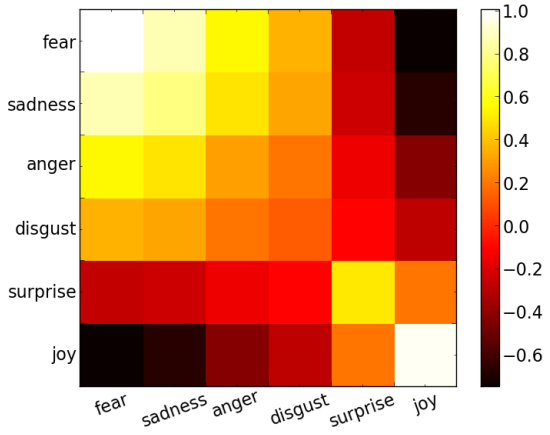

Figure 1: Heatmap showing a learned coregionalisation matrix over the emotions.

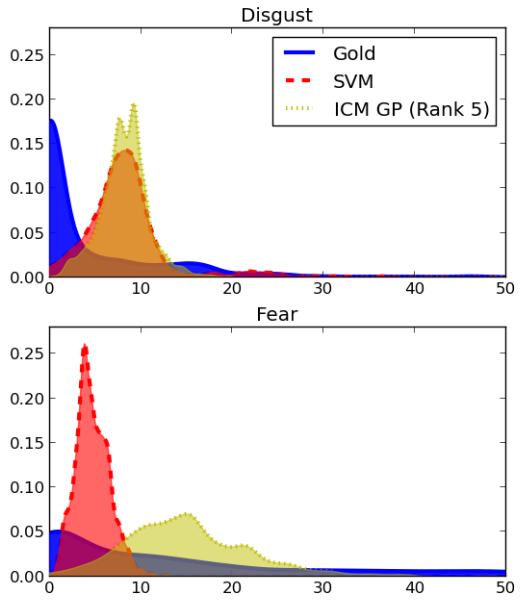

Figure 2: Test score distributions for *disgust* and *fear*. For clarity, only scores between 0 and 50 are shown. SVM performs better on *disgust*, while GP performs better on *fear*.

4.2 Prediction Results

Table 1 shows the Pearson's scores obtained in our experiments. The low-rank models outperformed the baselines for the full task (predicting all emotions) and for *fear*, *joy* and *surprise* subtasks. The rank 5 models were also able to outperform all GP baselines for the remaining emotions, but could not beat the SVM baseline. As expected, the "Combined" and "Combined+" performed worse than the low-rank models, probably due to their inability to model anti-correlations.

4.3 Error analysis

To check why SVM performs better than GPs for some emotions, we analysed their gold-standard score distributions. Figure 2 shows the smoothed distributions for *disgust* and *fear*, comparing the gold-standard scores to predictions from the SVM and GP models. The distributions for the training set follow similar shapes.

We can see that GP obtains better matching score distributions in the case when the gold-standard scores are more spread over the full support of response values, i.e., $[0, 100]$. Since our GP model employs a Gaussian likelihood, it is effectively minimising a squared-error loss. The SVM model, on the other hand, uses hinge loss, which is linear beyond the margin envelope constraints. This affects the treatment of outlier points, which attract quadratic cf. linear penalties for the GP and SVM respectively. Therefore, when training scores are more uniformly distributed (which is the case for *fear*), the GP model has to take the high scores into account, resulting in broader coverage of the full support. For *disgust*, the scores are much more peaked near zero, favouring the

more narrow coverage of the SVM.

More importantly, Figure 2 also shows that both SVM and GP predictions tend to exhibit a Gaussian shape, while the true scores show an exponential behaviour. This suggests that both models are making wrong prior assumptions about the underlying score distribution. For SVMs, this is a non-trivial issue to address, although it is much easier for GPs, where we can use a different likelihood distribution, e.g., a Beta distribution to reflect that the outputs are only valid over a bounded range. Note that non-Gaussian likelihoods mean that exact inference is no longer tractable, due to the lack of conjugacy between the prior and likelihood. However a number of approximate inference methods are appropriate which are already widely used in the GP literature for use with non-Gaussian likelihoods, including expectation propagation (Jylänki et al., 2011), the Laplace approximation (Williams and Barber, 1998) and Markov Chain Monte Carlo sampling (Adams et al., 2009).

4.4 Training Set Influence

We expect multi-task models to perform better for smaller datasets, when compared to single-task models. This stems from the fact that with small datasets often there is more uncertainty associated with each task, a problem which can be alleviated using statistics from the other tasks. To measure this behaviour, we performed an additional experiment varying the size of the training sets, while using 100 sentences for testing.

Figure 3 shows the scores obtained. As expected, for smaller datasets the single-task models are outperformed by ICM, but their performance become equivalent as the training set size increases. SVM performance tends to be slightly worse for most sizes. To study why we obtained an outlier for the single-task model with 200 sentences, we inspected the prediction values. We found that, in this case, predictions for *joy*, *surprise* and *disgust* were all around the same value.[6] For larger datasets, this effect disappears and the single-task models yield good predictions.

5 Conclusions and Future Work

This paper proposed an multi-task approach for Emotion Analysis that is able to learn correlations and anti-correlations between emotions. Our formulation is based on a combination of a Gaussian Process and a low-rank coregionalisation model, using a richer parameterisation that allows the learning of fine-grained task similarities. The proposed model outperformed strong baselines when applied to a news headline dataset.

As it was discussed in Section 4.3, we plan to further explore the possibility of using non-Gaussian likelihoods with the GP models. Another research avenue we intend to explore is to employ multiple layers of metadata, similar to the model proposed by Cohn and Specia (2013). An example is to incorporate the dataset provided by Snow et al. (2008), which provides multiple non-expert emotion annotations for each sentence, obtained via crowdsourcing. Finally, another possible extension comes from more advanced vector-valued GP models, such as the linear model of coregionalisation (Álvarez et al., 2012) or hierarchical kernels (Hensman et al., 2013). These models can be specially useful when we want to employ multiple kernels to explain the relation between the input data and the labels.

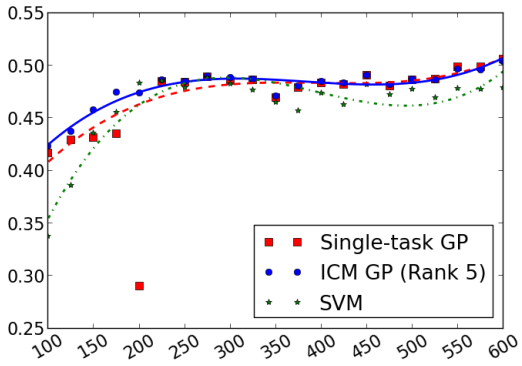

Figure 3: Pearson's correlation score according to training set size (in number of sentences).

Acknowledgements

Daniel Beck was supported by funding from CNPq/Brazil (No. 237999/2012-9). Dr. Cohn is the recipient of an Australian Research Council Future Fellowship (project number FT130101105).

References

Ryan Prescott Adams, Iain Murray, and David J. C. MacKay. 2009. Tractable Nonparametric Bayesian

[6] Looking at the predictions for smaller datasets, we found the same behaviour, but because the values found were near the mean they did not hurt the Pearson's score as much.

Inference in Poisson Processes with Gaussian Process Intensities. In *Proceedings of ICML*, pages 1–8, New York, New York, USA. ACM Press.

Mauricio A. Álvarez, Lorenzo Rosasco, and Neil D. Lawrence. 2012. Kernels for Vector-Valued Functions: a Review. *Foundations and Trends in Machine Learning*, pages 1–37.

Daniel Beck, Kashif Shah, Trevor Cohn, and Lucia Specia. 2013. SHEF-Lite : When Less is More for Translation Quality Estimation. In *Proceedings of WMT13*, pages 337–342.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Edwin V. Bonilla, Kian Ming A. Chai, and Christopher K. I. Williams. 2008. Multi-task Gaussian Process Prediction. *Advances in Neural Information Processing Systems*.

Rich Caruana. 1997. Multitask Learning. *Machine Learning*, 28:41–75.

Trevor Cohn and Lucia Specia. 2013. Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation. In *Proceedings of ACL*.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*.

Jenny Rose Finkel and Christopher D. Manning. 2009. Hierarchical Bayesian Domain Adaptation. In *Proceedings of NAACL*.

James Hensman, Neil D Lawrence, and Magnus Rattray. 2013. Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinformatics*, 14:252.

Pasi Jylänki, Jarno Vanhatalo, and Aki Vehtari. 2011. Robust Gaussian Process Regression with a Student-t Likelihood. *Journal of Machine Learning Research*, 12:3227–3257.

Rada Mihalcea and Carlo Strapparava. 2012. Lyrics, Music, and Emotions. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 590–599.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Duborg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Daniel Preotiuc-Pietro and Trevor Cohn. 2013. A temporal model of text periodicities using Gaussian Processes. In *Proceedings of EMNLP*.

Carl Edward Rasmussen and Christopher K. I. Williams. 2006. *Gaussian processes for machine learning*, volume 1. MIT Press Cambridge.

Kashif Shah, Trevor Cohn, and Lucia Specia. 2013. An Investigation on the Effectiveness of Features for Translation Quality Estimation. In *Proceedings of MT Summit XIV*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and Fast - But is it Good?: Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of EMNLP*.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14 : Affective Text. In *Proceedings of SEMEVAL*.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*.

Christopher K. I. Williams and David Barber. 1998. Bayesian Classification with Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351.