Predictive Automatic Relevance Determination by Expectation Propagation

Alan Qi
Thomas P. Minka
Rosalind W. Picard
Zoubin Ghahramani
Motivation

Task 1: Classify high dimensional datasets with many irrelevant features e.g., normal v.s. cancer microarray data.

Task 2: Sparse Bayesian kernel classifiers for fast test performance
Bayesian Classification Model

Prior of the classifier $w$:

$$p(w | \alpha) = \prod_i \mathcal{N}(w_i | 0, \alpha_i^{-1})$$

Likelihood for the data:

$$p(t | w, X) = \prod_i p(t_i | x_i, w) = \prod_i \Psi(t_i w^T \phi(x_i))$$

Where $t = \{t_i\}_{i=1}^N$, $X = \{x_i\}_{i=1}^N$, $\Psi(\cdot)$ is a cumulative distribution function for a standard Guassian.
Evidence and Predictive distribution

The evidence, i.e., the marginal likelihood of the hyperparameters $\alpha$:

$$p(t|\alpha) = \int p(w, t|\alpha) dw$$

The predictive posterior distribution of the label $t_{N+1}$ for a new input $x_{N+1}$:

$$p(t_{N+1}|x_{N+1}, t) = \int p(t_{N+1}|x_{N+1}, w)p(w|t) dw$$
Automatic Relevance Determination (ARD)

• Give the feature weights independent Gaussian priors whose variance, $\alpha^{-1}$, controls how far away from zero each weight is allowed to go.

• Maximize $p(t | \alpha)$, the marginal likelihood of the model with respect to $\alpha$.

• Outcome: many elements of $\alpha$ go to infinity, which naturally prunes irrelevant features in the data.
Risk of Optimizing $\alpha$

- Max-Margin
- Max-Evd-1
- Bayes Point
- Max-Evd-2

X: Class 1 vs O: Class 2
Two types of overfitting

• Classical Maximum likelihood:
  – Optimizing the classifier weights $w$ can directly fit noise in the data, resulting a complicated model.

• Type II Maximum likelihood ($\text{ARD}_\alpha$):
  – Optimizing the hyperparameters corresponds to choose which variables are irrelevant. Choosing a simple model can also overfit if we maximize the model marginal likelihood.
Overfitting by Type II ML training

• Particularly, if maximizing the marginal likelihood of the model and the dimension of (the number of the features) is large, there exists the risk of overfitting (as shown in some practical examples).
Predictive-ARD

- Choosing the model with the best estimated *predictive performance* instead of the most probable model.
- Expectation propagation (EP) estimates the leave-one-out predictive performance without performing any expensive cross-validation.
Estimate Predictive Performance

• Predictive posterior given a test data point $\mathbf{x}_{N+1}$
  \[ p(t_{N+1} \mid \mathbf{x}_{N+1}, D) = \int p(t_{N+1} \mid \mathbf{x}_{N+1}, \mathbf{w}) p(\mathbf{w} \mid D) d\mathbf{w} \]

• EP estimate of predictive leave-one-out error probability
  \[ \frac{1}{N} \sum_{i=1}^{N} (1 - p(t_i \mid \mathbf{x}_i, D_{\setminus i})) \approx \frac{1}{N} \sum_{i=1}^{N} \left(1 - \int p(t_i \mid \mathbf{x}_i, \mathbf{w}) q(\mathbf{w} \mid D_{\setminus i}) d\mathbf{w} \right) \]

• EP estimate of predictive leave-one-out error count
  \[ LOO = \frac{1}{N} \sum_{i=1}^{N} I( p(t_i \mid \mathbf{x}_i, D_{\setminus i}) < \frac{1}{2} ) \]
Expectation Propagation in a Nutshell

- Approximate a probability distribution by simpler parametric terms:

\[
p(w | t) = \prod_i f_i(w) = \prod_i \Psi(t_i w^T \phi(x_i))
\]

\[
q(w) = \prod_i \tilde{f}_i(w)
\]

- Each approximation term \( \tilde{f}_i(w) \) lives in an exponential family (e.g. Gaussian)
EP in a Nutshell

Three key steps:

• Deletion Step: approximate the “leave-one-out” predictive posterior for the $i^{th}$ point:
  \[ q^i(w) \propto q(w) / \tilde{f}_i(w) \]

• Minimizing the following KL divergence by moment matching:
  \[ \arg\min_{\tilde{f}_i(x)} D(f_i(x)q^i(x) \| \tilde{f}_i(x)q^i(x)) \]

• Inclusion: \[ q^i(x) = \tilde{f}_i(x)q^i(x) \]

The key observation: we can use the approximate predictive posterior, obtained in deletion step, for model selection. No extra computation!
Sequential Update

• EP approximates true likelihood terms by Gaussian virtual observations.
• Based on Gaussian virtual observations, the classification model becomes a regression model.
• Then, we can achieve efficient sequential updates without maintaining and updating a full covariance matrix. (Faul & Tipping 02)
Comparison of different model selection criteria for ARD training

1st row: Test error  
2nd row: Estimated leave-one-out error probability  
3rd row: Estimated leave-one-out error counts  
4th row: Evidence (Model marginal likelihood)  
5th row: Fraction of selected features

(a) Synthetic data classification  
(b) Leukaemia data classification
Gene Expression Classification

Task: Classify gene expression datasets into different categories, e.g., normal v.s. cancer

Challenge: Thousands of genes measured in the micro-array data. Only a small subset of genes are probably correlated with the classification task.
Classifying Leukemia Data

- The task: distinguish acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL).
- The dataset: 47 and 25 samples of type ALL and AML respectively with 7129 features per sample.
- The dataset was randomly split 100 times into 36 training and 36 testing samples.
Classifying Colon Cancer Data

- The task: distinguish normal and cancer samples.
- The dataset: 22 normal and 40 cancer samples with 2000 features per sample.
- The dataset was randomly split 100 times into 50 training and 12 testing samples.
- SVM results from Li et al. 2002
Bayesian Sparse Kernel Classifiers

• Using feature/kernel expansions defined on training data points:
\[ \phi(x_i) = [1, k(x_i, x_1), \ldots, k(x_i, x_N)]^T \]

• Predictive-ARD-EP trains a classifier that depends on a small subset of the training set.

• Fast test performance.
Test error rates and numbers of relevance or support vectors on breast cancer dataset.

50 partitionings of the data were used. All these methods use the same Gaussian kernel with kernel width $= 5$. The trade-off parameter $C$ in SVM is chosen via 10-fold cross-validation for each partition.
Test error rates on diabetes data

100 partitionings of the data were used. Evidence and Predictive ARD-EPs use the Gaussian kernel with kernel width = 5.
Summary

• ARD is an excellent Bayesian feature selection and sparse learning method.
• However, maximizing marginal likelihood can lead to overfitting if there are a lot of features.
• We propose Predictive ARD based on EP
• In practice it works very well.
• Related work: Opper & Winther 2000.