Machine Learning 2

7. Sparse Linear Models – Further Methods

Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)
Institute for Computer Science
University of Hildesheim, Germany

Outline

1. Automatic Relevance Determination (ARD)

Syllabus

		A. Advanced Supervised Learning
Tue. 9.12.	(1)	A.1 Generalized Linear Models
Wed. 10.12.	(2)	A.2 Gaussian Processes
Tue. 16.12.	(3)	A.3 Advanced Support Vector Machines
Wed. 17.12.	(4)	A.4 Neural Networks
Tue. 6.1.	(5)	A.5 Ensembles
Wed. 7.1.	(6)	A.5b Ensembles (ctd.)
Tue. 13.1.	(7)	A.6 Sparse Linear Models — L1 regularization
Wed. 14.1.	(8)	A.6b Sparse Linear Models — L1 regularization (ctd.)
Tue. 20.1.	(9)	A.7. Sparse Linear Models — Further Methods
Wed. 21.1.	(10)	
Tue. 27.1.	(11)	
Wed. 28.1.	(12)	
Tue. 3.2.	(13)	
Wed. 4.2.	(14)	

Outline

1. Automatic Relevance Determination (ARD)

Linear Regression plus ARD Regularization

Linear Regression plus L2 regularization (Ridge Regression):

$$p(y_n \mid x_n, \beta, \sigma_y^2) := \mathcal{N}(y_n \mid \beta^T x_n, \sigma_y^2)$$
$$p(\beta) := \mathcal{N}(\beta \mid 0, \Sigma_\beta := \sigma_\beta^2 I)$$

Linear Regression plus ARD Regularization:

$$p(y_n \mid x_n, \beta, \sigma_y^2) := \mathcal{N}(y_n \mid \beta^T x_n, \sigma_y^2)$$
$$p(\beta) := \mathcal{N}(\beta \mid 0, \Sigma_\beta := \mathrm{diag}(\sigma_{\beta_1}^2, \ldots, \sigma_{\beta_M}^2))$$

Linear Regression plus ARD Regularization

Linear Regression plus L2 regularization (Ridge Regression):

$$p(y_n \mid x_n, \beta, \sigma_y^2) := \mathcal{N}(y_n \mid \beta^T x_n, \sigma_y^2)$$
$$p(\beta) := \mathcal{N}(\beta \mid 0, \Sigma_\beta := \sigma_\beta^2 I)$$

Linear Regression plus ARD Regularization:

$$p(y_n \mid x_n, \beta, \sigma_y^2) := \mathcal{N}(y_n \mid \beta^T x_n, \sigma_y^2)$$
$$p(\beta) := \mathcal{N}(\beta \mid 0, \Sigma_\beta := \mathrm{diag}(\sigma_{\beta_1}^2, \ldots, \sigma_{\beta_M}^2))$$

Idea:

- use a different regularization weight for each predictor x_m.

Linear Regression plus ARD Regularization

Linear Regression plus L2 regularization (Ridge Regression):

$$p(y_n \mid x_n, \beta, \sigma_y^2) := \mathcal{N}(y_n \mid \beta^T x_n, \sigma_y^2)$$
$$p(\beta) := \mathcal{N}(\beta \mid 0, \Sigma_\beta := \sigma_\beta^2 I)$$

Linear Regression plus ARD Regularization:

$$p(y_n \mid x_n, \beta, \sigma_y^2) := \mathcal{N}(y_n \mid \beta^T x_n, \sigma_y^2)$$
$$p(\beta) := \mathcal{N}(\beta \mid 0, \Sigma_\beta := \mathrm{diag}(\sigma_{\beta_1}^2, \ldots, \sigma_{\beta_M}^2))$$

Idea:

- use a different regularization weight for each predictor x_m.
- but M hyperparameters are too many to learn by grid search.

Linear Regression plus ARD Regularization

Linear Regression plus L2 regularization (Ridge Regression):

$$p(y_n \mid x_n, \beta, \sigma_y^2) := \mathcal{N}(y_n \mid \beta^T x_n, \sigma_y^2)$$
$$p(\beta) := \mathcal{N}(\beta \mid 0, \Sigma_\beta := \sigma_\beta^2 I)$$

Linear Regression plus ARD Regularization:

$$p(y_n \mid x_n, \beta, \sigma_y^2) := \mathcal{N}(y_n \mid \beta^T x_n, \sigma_y^2)$$
$$p(\beta) := \mathcal{N}(\beta \mid 0, \Sigma_\beta := \mathrm{diag}(\sigma_{\beta_1}^2, \ldots, \sigma_{\beta_M}^2))$$
$$p(\sigma_y^2) := \mathrm{IG}(\sigma_y^2 \mid c, d)$$
$$p(\sigma_{\beta_m}^2) := \mathrm{IG}(\sigma_{\beta_m}^2 \mid a, b), \quad m = 1, \ldots, M$$

Idea:

- use a different regularization weight for each predictor x_m.
- but M hyperparameters are too many to learn by grid search.
- hence put a hyperprior on top.

Empirical Bayes

Maximum Likelihood (ML): $\hat{\theta} := \arg\max_{\theta} p(\mathcal{D} \mid \theta)$

Full Bayes: $(\hat{\theta}, \hat{\eta}) \sim p(\theta, \eta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)\, p(\theta \mid \eta)\, p(\eta)$

Empirical Bayes

Maximum Likelihood (ML): $\hat{\theta} := \arg\max_{\theta} p(\mathcal{D} \mid \theta)$

Maximum Aposteriori (MAP): $\hat{\theta} := \arg\max_{\theta} p(\mathcal{D} \mid \theta)\, p(\theta \mid \eta)$

Full Bayes: $(\hat{\theta}, \hat{\eta}) \sim p(\theta, \eta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)\, p(\theta \mid \eta)\, p(\eta)$

Empirical Bayes

Maximum Likelihood (ML): $\hat{\theta} := \arg\max_{\theta} p(\mathcal{D} \mid \theta)$

Maximum Aposteriori (MAP): $\hat{\theta} := \arg\max_{\theta} p(\mathcal{D} \mid \theta)\, p(\theta \mid \eta)$

ML-II (Empirical Bayes):

$$\hat{\eta} := \arg\max_{\eta} p(\mathcal{D} \mid \eta)$$

$$= \arg\max_{\eta} \int p(\mathcal{D} \mid \theta)\, p(\theta \mid \eta) d\theta$$

$$\hat{\theta} \sim p(\theta \mid \mathcal{D}, \hat{\eta}) \propto p(\mathcal{D} \mid \theta)\, p(\theta \mid \hat{\eta})$$

MAP-II:

$$\hat{\eta} := \arg\max_{\eta} p(\mathcal{D} \mid \eta)\, p(\eta)$$

$$= \arg\max_{\eta} \int p(\mathcal{D} \mid \theta)\, p(\theta \mid \eta)\, p(\eta) d\theta$$

$$\hat{\theta} \sim p(\theta \mid \mathcal{D}, \hat{\eta}) \propto p(\mathcal{D} \mid \theta)\, p(\theta \mid \hat{\eta})$$

Full Bayes: $(\hat{\theta}, \hat{\eta}) \sim p(\theta, \eta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)\, p(\theta \mid \eta)\, p(\eta)$

Marginal Likelihood

Without hyperpriors:

$$p(y \mid X, \sigma_y^2, \Sigma_\beta) = \int \mathcal{N}(y \mid X\beta, \sigma_y^2 I)\mathcal{N}(\beta \mid 0, \Sigma_\beta)d\beta$$
$$= \mathcal{N}(y \mid 0, \sigma_y^2 I + X\Sigma_\beta X^T)$$

Marginal Likelihood

Without hyperpriors:

$$p(y \mid X, \sigma_y^2, \Sigma_\beta) = \int \mathcal{N}(y \mid X\beta, \sigma_y^2 I)\mathcal{N}(\beta \mid 0, \Sigma_\beta) d\beta$$
$$= \mathcal{N}(y \mid 0, \underbrace{\sigma_y^2 I + X\Sigma_\beta X^T}_{=:C_y})$$

$$\ell(\sigma_y^2, \Sigma_\beta) := \log p(y \mid X, \sigma_y^2, \Sigma_\beta) \propto \log|C_y| + y^T C_y^{-1} y$$

Marginal Likelihood

Without hyperpriors:

$$p(y \mid X, \sigma_y^2, \Sigma_\beta) = \int \mathcal{N}(y \mid X\beta, \sigma_y^2 I)\mathcal{N}(\beta \mid 0, \Sigma_\beta)d\beta$$
$$= \mathcal{N}(y \mid 0, \underbrace{\sigma_y^2 I + X\Sigma_\beta X^T}_{=:C_y})$$

$$\ell(\sigma_y^2, \Sigma_\beta) := \log p(y \mid X, \sigma_y^2, \Sigma_\beta) \propto \log|C_y| + y^T C_y^{-1} y$$

With hyperpriors:

$$\ell(\sigma_y^2, \Sigma_\beta) := \log p(y \mid X, \sigma_y^2, \Sigma_\beta, a, b, c, d)$$
$$\propto \log|C_y| + y^T C_y^{-1} y + \sum_{m=1}^{M}(-a \log \sigma_{\beta_m}^2 - b/\sigma_{\beta_m}^2)$$
$$- c \log \sigma_y^2 - d/\sigma_y^2$$

Inferring Parameters β

$$p(\beta \mid X, y, \sigma_y^2, \Sigma_\beta) = \frac{1}{Z}\mathcal{N}(\beta \mid 0, \Sigma_\beta)\mathcal{N}(y \mid X\beta, \sigma_y^2 I)$$
$$= \mathcal{N}(\beta \mid \mu_\beta := \frac{1}{\sigma_y^2} C_\beta X^T y, C_\beta := (\frac{1}{\sigma_y^2} X^T X + \Sigma_\beta^{-1})^{-1})$$

for $\Sigma_\beta = \infty I$: unregularized estimates

$$= \mathcal{N}(\beta \mid (X^T X)^{-1} X^T y, \sigma_y^2 (X^T X)^{-1})$$

for $\sigma_y^2 = \infty$: overregularized estimates

$$= \mathcal{N}(\beta \mid 0, \Sigma_\beta)$$

Equivalent MAP Estimation Problem

$$\ell(\beta) = \frac{1}{\sigma_y^2}||y - X\beta||_2^2 + \min_{\sigma^2_{\beta_1},\ldots,\sigma^2_{\beta_M} \geq 0} \log|\sigma_y^2 I + X\mathrm{diag}(\sigma_\beta^2)X^T| + \sum_{m=1}^{M} \frac{\beta_m^2}{\sigma^2_{\beta_m}}$$

Learning ARD I: via EM

$$\ell(\sigma_y^2, \Sigma_\beta, \mu_\beta, C_\beta) := E_{\beta \sim \mathcal{N}(\mu_\beta, C_\beta)}(\log p(y \mid X, \beta, \sigma_y^2, \Sigma_\beta))$$

$$= E_{\beta \sim \mathcal{N}(\mu_\beta, C_\beta)}(\log \mathcal{N}(y \mid X\beta, \sigma_y^2) + \log \mathcal{N}(\beta \mid 0, \Sigma_\beta))$$

$$+ \log \mathrm{IG}(\sigma_y^2 \mid a, b) + \sum_{m=1}^{M} \log \mathrm{IG}(\sigma_{\beta_m}^2 \mid c, d)$$

$$= E_{\beta \sim \mathcal{N}(\mu_\beta, C_\beta)}(-\frac{N}{2} \log \sigma_y^2 - \frac{1}{2\sigma_y^2} ||y - X\beta||^2 - \frac{1}{2} \sum_m \log \sigma_{\beta_m}^2 - \frac{1}{2} \mathrm{tr} \Sigma_\beta^{-1} \beta \beta^T$$

$$+ \sum_{m=1}^{M} (-a \log \sigma_{\beta_m}^2 - b/\sigma_{\beta_m}^2) - c \log \sigma_y^2 - d/\sigma_y^2$$

$$= -\frac{N}{2} \log \sigma_y^2 - \frac{1}{2\sigma_y^2}(||y - X\mu_\beta||^2 + \mathrm{tr}(X^T X C_\beta)) - \frac{1}{2} \sum_m \log \sigma_{\beta_m}^2$$

$$- \frac{1}{2} \mathrm{tr} \Sigma_\beta^{-1} (\mu_\beta \mu_\beta^T + C_\beta) + \sum_{m=1}^{M} (-a \log \sigma_{\beta_m}^2 - b/\sigma_{\beta_m}^2) - c \log \sigma_y^2 - d/\sigma_y^2$$

Learning ARD I: via EM

$$
\begin{aligned}
&\ell(\ldots) \\
&= -\frac{N}{2}\log\sigma_y^2 - \frac{1}{2\sigma_y^2}(||y - X\mu_\beta||^2 + \mathrm{tr}(X^T X C_\beta)) - \frac{1}{2}\sum_m \log\sigma_{\beta_m}^2 \\
&- \frac{1}{2}\mathrm{tr}\Sigma_\beta^{-1}(\mu_\beta\mu_\beta^T + C_\beta) + \sum_{m=1}^{M}(-a\log\sigma_{\beta_m}^2 - b/\sigma_{\beta_m}^2) - c\log\sigma_y^2 - d/\sigma_y^2 \\
&\propto -(2c+N)\log\sigma_y^2 - (2d + ||y - X\mu_\beta||^2 + \mathrm{tr}(X^T X C_\beta))\frac{1}{\sigma_y^2} \\
&\quad - \sum_m (2a+1)\log\sigma_{\beta_m}^2 + 2b/\sigma_{\beta_m}^2 - \mathrm{tr}\Sigma_\beta^{-1}(\mu_\beta\mu_\beta^T + C_\beta)
\end{aligned}
$$

Learning ARD I: via EM

$$\ell(\ldots) = -(2c+N)\log\sigma_y^2 - (2d + ||y - X\mu_\beta||^2 + \mathrm{tr}(X^T X C_\beta))\frac{1}{\sigma_y^2}$$
$$- \sum_m (2a+1)\log\sigma_{\beta_m}^2 + 2b/\sigma_{\beta_m}^2 - \mathrm{tr}\Sigma_\beta^{-1}(\mu_\beta\mu_\beta^T + C_\beta)$$

$$0 \stackrel{!}{=} \frac{\partial\ell}{\partial\sigma_{\beta_m}^2} = -(2a+1)\frac{1}{\sigma_{\beta_m}^2} + (2b + (\mu_\beta)_m^2 + (C_\beta)_{m,m})/(\sigma_{\beta_m}^2)^2$$

$$\sigma_{\beta_m}^2 = \frac{2b + (\mu_\beta)_m^2 + (C_\beta)_{m,m}}{2a+1}$$

$$0 \stackrel{!}{=} \frac{\partial\ell}{\partial\sigma_y^2}$$

$$\rightsquigarrow \sigma_y^2 = \frac{2d + ||y - X\mu_\beta||^2 + \mathrm{tr}(X^T X C_\beta)}{2c+N}$$

which can be accelerated using

$$C_\beta X^T X = \sigma_y^{2\mathrm{old}}(I - C_\beta\Sigma_\beta^{-1}), \quad \mathrm{tr}(\ldots) = \sigma_y^{2\mathrm{old}}\sum 1 - \frac{(C_\beta)_{m,m}}{\sigma_\beta^2}$$

Learning ARD II: Fixed Point Algorithm

iteratively fit:

$$\sigma^2_{\beta_m} := \frac{2b + (\mu_\beta)^2_m}{2a + \gamma_m}$$

$$\sigma^2_y := \frac{2d + ||y - X\mu_\beta||^2}{2c + N - \sum_m \gamma_m}$$

$$C_\beta := (\frac{1}{\sigma^2_y} X^T X + \Sigma_\beta^{-1})^{-1}$$

$$\mu_\beta := \frac{1}{\sigma^2_y} C_\beta X^T y$$

$$\gamma_m := 1 - \frac{(C_\beta)_{m,m}}{\sigma^2_{\beta_m}}, \quad m := 1, \ldots, M$$

Learning ARD III: Iteratively Reweighted L1

The ARD regularization term

$$R(\sigma_\beta^2) := \log|C_y(\sigma_\beta^2)| = \log|\sigma_y^2 I + X\Sigma_\beta X^T|, \quad \Sigma_\beta := \mathrm{diag}(\sigma_\beta^2)$$

is concave in σ_β^2 and thus can be written as

$$R(\sigma_\beta^2) = \min_\lambda \lambda^T \sigma_\beta^2 - R^*(\lambda)$$

$$R^*(\lambda) = \min_{\tilde{\sigma}_\beta^2} \lambda^T \tilde{\sigma}_\beta^2 - \log|C_y(\tilde{\sigma}_\beta^2)|$$

The relaxed function

$$R(\sigma_\beta^2, \lambda) := \lambda^T \sigma_\beta^2 - R^*(\lambda) = \lambda^T \sigma_\beta^2 - \min_{\tilde{\sigma}_\beta^2} \lambda^T \tilde{\sigma}_\beta^2 - \log|C_y(\tilde{\sigma}_\beta^2)|$$

for fixed σ_β^2 is minimized by

$$\lambda = \nabla_{\sigma_\beta^2} \log|C_y(\sigma_\beta^2)|$$

Learning ARD III: Iteratively Reweighted L1

Instead of σ^2_β Wipf/Nagarajan 2008 use

$$\sigma^2_{\beta_m} \overset{??}{\rightarrow} \lambda_m^{\frac{1}{2}} |\beta_m|$$

finally yielding the iterative procedure:

$$\beta^{(t+1)} := \arg\min_\beta \ell(\beta) + \sum_{m=1}^{M} \lambda_m^{(t)} |\beta_m|$$

and to find $\lambda^{(t)}$:

$$\lambda_m^{(0)} := 1$$

$$\lambda_m^{(t+1)} := (X_{.,m}(\sigma_y^2 I + X \mathrm{diag}(\frac{1}{\lambda_1^{(t)}}, \ldots, \frac{1}{\lambda_M^{(t)}}) \mathrm{diag}(|\beta_1^{(t)}|, \ldots, |\beta_M^{(t)}|))^{-1} X_{.,m})^{\frac{1}{2}}$$

ARD for Classification

- so far, everything was developed for linear regression.
- for logistic regression, for EM the E-step cannot be done analytically.
  - possibly use variational approximation
  - use Gaussian approximation (Laplace approximation)
- the iteratively reweighted learning algorithm still works.

Remarks

- ARD is a good example for a (arguably simple) hierarchical Bayesian model.
- ARD has to be diligently evaluated against simple baselines such as normalizing the data with a vanilla L1/L2 regularized model.

Further Readings

- L1 regularization: [Mur12, chapter 13.3–5], [HTFF05, chapter 3.4, 3.8, 4.4.4], [Bis06, chapter 3.1.4].
  - LAR, LARS: [HTFF05, chapter 3.4.4], [Mur12, chapter 13.4.2],
- Non-convex regularizers: [Mur12, chapter 13.6].
- Automatic Relevance Determination (ARD): [Mur12, chapter 13.7], [HTFF05, chapter 11.9.1], [Bis06, chapter 7.2.2].
- Sparse Coding: [Mur12, chapter 13.8].

References

Christopher M. Bishop.
Pattern recognition and machine learning, volume 1.
springer New York, 2006.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin.
The elements of statistical learning: data mining, inference and prediction, volume 27.
Springer, 2005.

Kevin P. Murphy.
Machine learning: a probabilistic perspective.
The MIT Press, 2012.