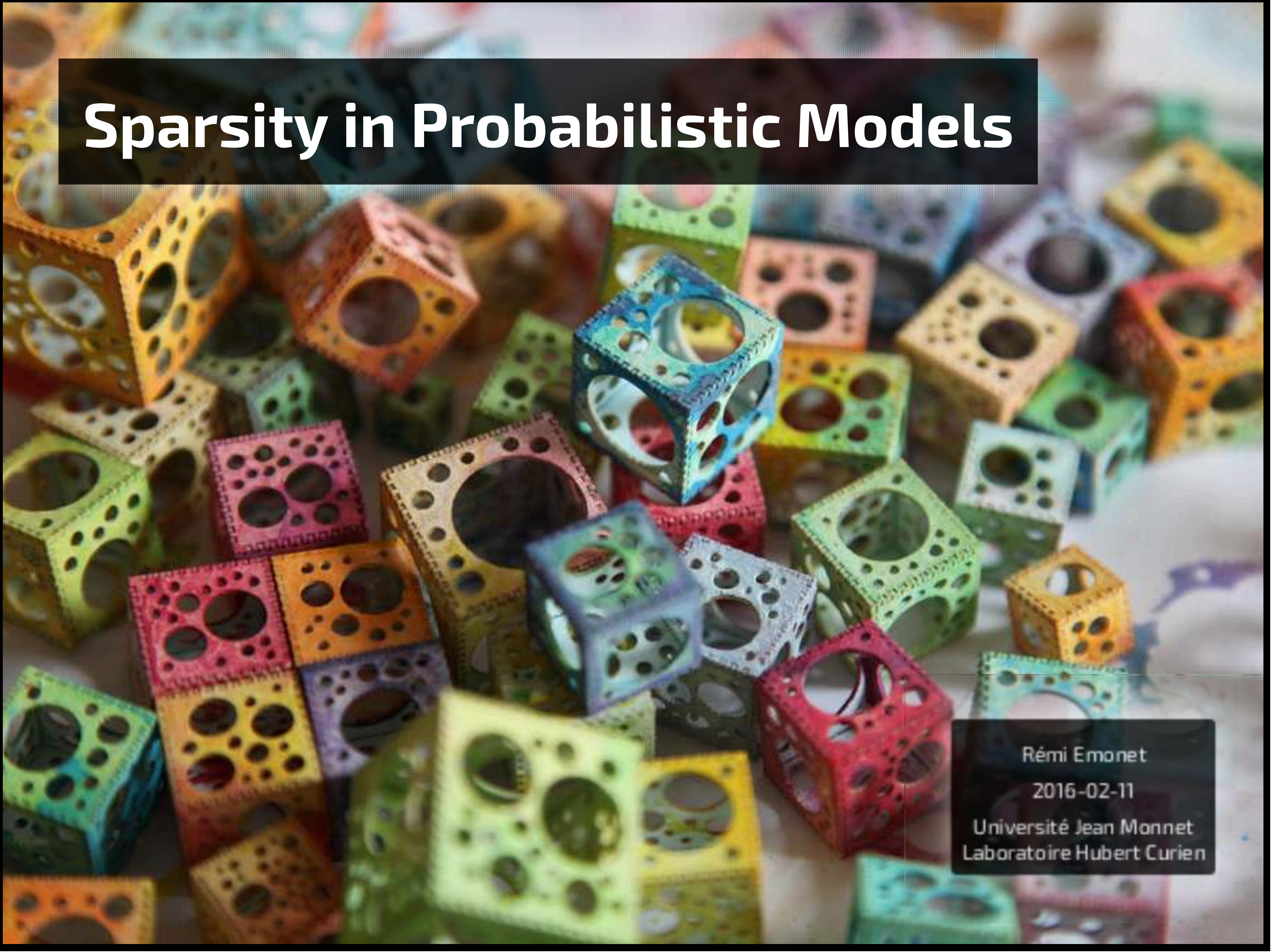


Sparsity in Probabilistic Models

Rémi Emonet

2016-02-11

Université Jean Monnet
Laboratoire Hubert Curien

Sparsity in Probabilistic Models

- Least Square Linear Regression:
  a Probabilistic View
- Regularized Least Square Regression:
  a Bayesian View
- Mixture Models and Alternatives to Model Selection
- Learning Sparse Matrices
- Conclusion

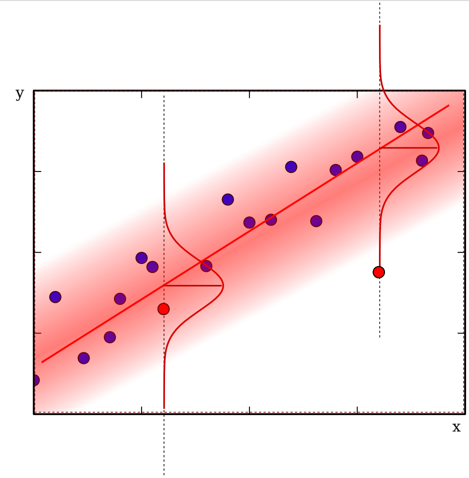
UNIVERSITÉ JEAN MONNET SAINT-ÉTIENNE

LABORATOIRE HUBERT CURIEN UMR • CNRS • 5516 • SAINT-ETIENNE

Linear Regression: a Probabilistic View

Linear Regression: a Probabilistic Model

- We observe $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
- We model $y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$,
  with $\epsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$
  - equivalent to: $y_i \sim \mathrm{N}\left(\mathbf{w}^T \mathbf{x}_i, \sigma^2\right)$
  - NB: as the noise was independent, y_i are independent, given $\mathbf{w}$



- The parameters of the model are $\mathbf{w}$
- The likelihood is $L(\mathbf{w}, S) = p(S|\mathbf{w})$
  - from the Independence $L(\mathbf{w}, S) = \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{y}_i | \mathbf{w})$
  - from the Normal distribution: $L(\mathbf{w}, S) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(y_i - \mathbf{w}^T \mathbf{x}_i)^2}{2\sigma^2}}$

Linear Regression: maximum likelihood

- Reminder
    - likelihood: $L(\mathbf{w}, S) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2}\right)$
- We want to maximize the likelihood, over $\mathbf{w}$,
    - we will rather consider the log-likelihood
        - $\log L(\mathbf{w}, S) = \sum_{i=1}^{n} \left(-\log(\sigma\sqrt{2\pi}) + \frac{-(y_i - \mathbf{w}^T\mathbf{x}_i)^2}{2\sigma^2}\right)$
        - $\log L(\mathbf{w}, S) = -n\log(\sigma\sqrt{2\pi}) + \frac{1}{2\sigma^2}\sum_{i=1}^{n} -(y_i - \mathbf{w}^T\mathbf{x}_i)^2$
    - we have
        - $\arg\max_w \; L(\mathbf{w}, S) = \arg\max_w \; \log L(\mathbf{w}, S)$
        - $\arg\max_w \; L(\mathbf{w}, S) = \arg\max_w \; \frac{1}{2\sigma^2}\sum_{i=1}^{n} -(y_i - \mathbf{w}^T\mathbf{x}_i)^2$
        - $\arg\max_w \; L(\mathbf{w}, S) = \arg\min_w \; \sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i)^2$

Linear Regression: summary

- Supposing a Gaussian noise around the linear predictor $\mathbf{w}$
- These are equivalent point of views to find $\mathbf{w}$
  - maximizing the likelihood $L(\mathbf{w}, S) = \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{y}_i | \mathbf{w})$
  - minimizing $\frac{1}{2} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2$
  - solving the least square linear regression problem
- NB: the noise variance σ^2 does not appear, cool!?

Regularized Least Square: a Bayesian View

Should we really maximize the likelihood?

- We have observations (S) and

  we want to find the best parameters $\mathbf{w}$

- We actually want to find
  - the parameters that are the most supported by the observations
  - i.e., the parameters that are most likely, knowing the observations
  - i.e., $\arg\max_w \; p(\mathbf{w}|S)$ (reminder: the likelihood is $L(\mathbf{w}, S) = p(S|\mathbf{w})$)

- How is it related?
  - Bayes: $p(A|B)p(B) = p(A \wedge B) = p(B \wedge A) = p(B|A)p(A)$
  - Bayes, v2: $p(A|B) = \dfrac{p(B|A)p(A)}{p(B)}$
  - so: $p(\mathbf{w}|S) = \dfrac{p(S|\mathbf{w})p(\mathbf{w})}{p(S)}$

Bayesian Posterior Optimization

- We have observations S, we want the best parameters $\mathbf{w}$
- We want to
  - maximize: $p(\mathbf{w}|S) = \dfrac{p(S|\mathbf{w})p(\mathbf{w})}{p(S)}$
  - i.e., maximize: $p(S|\mathbf{w})p(\mathbf{w})$ (as $p(S)$ does not depend en $\mathbf{w}$)
  - i.e, minimize: $-\log(p(S|\mathbf{w})p(\mathbf{w}))$ (as $\log$ is increasing)
  - i.e, minimize: $-\log(p(S|\mathbf{w})) - \log(p(\mathbf{w}))$
  - i.e, minimize: $-\log L(\mathbf{w}, S) - \log(p(\mathbf{w}))$
- One (possible) interpretation
  - we want to optimize the (negative-log)-likelihood (as in MLE) but **penalized** by the (negative-log)-prior

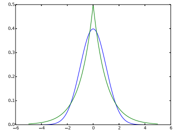
And Ockham?
(aside on the board)

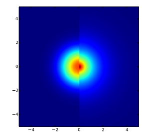
From Prior to Regularization

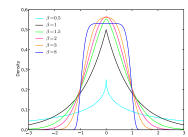
- Bayesian opt., minimize: $-\log L(\mathbf{w}, S) - \log(p(\mathbf{w}))$
- Back to a Gaussian noise σ^2 around the linear predictor
  - minimize: $\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 - \log(p(\mathbf{w}))$
  - i.e., minimize: $\frac{1}{2} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 - \sigma^2 \cdot \log(p(\mathbf{w}))$
- We can identify
  - the regularization for regularized the least square
  - and, $-\sigma^2 \cdot \log(p(\mathbf{w}))$ (the obs. noise variance times the log-prior)
  - NB: $\log p$ is negative, so it is really a penalty

Priors and (some) L_p Norms

- Regularization is identified to $-\sigma^2 \cdot \log(p(\mathbf{w}))$
- Isotropic Normal prior, i.e.: $\mathbf{w} \sim \mathrm{N}\left(0, \sigma_w^2 \mathbf{I}\right)$
  - $\log(p(\mathbf{w})) = cst + \log\exp\left(-\frac{\mathbf{w}^T\mathbf{w}}{2\sigma_w^2}\right)$
  - i.e., $-\log(p(\mathbf{w})) = \frac{\mathbf{w}^T\mathbf{w}}{2\sigma_w^2} - cst = \frac{\|\mathbf{w}\|_2^2}{2\sigma_w^2} - cst$
  - Regularizer: $\frac{\sigma^2}{2\sigma_w^2} \|\mathbf{w}\|_2^2$
- Laplace prior, i.e.: $\mathbf{w_j} \sim \mathrm{L}aplace\left(0, b_w\right)$
  - $\log(p(\mathbf{w})) = cst + \sum_j \log\exp\left(-\frac{|\mathbf{w_j}|}{b_w}\right)$
  - Regularizer: $\frac{\sigma^2}{b_w} \|\mathbf{w}\|_1$
- Generalized Normal distr. (v1): $p(x|\mu, \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left(-\frac{|x-\mu|^\beta}{\alpha^\beta}\right)$
  - Regularizer: $\frac{\sigma^2}{\alpha^\beta} \|\mathbf{w}\|_\beta^\beta$



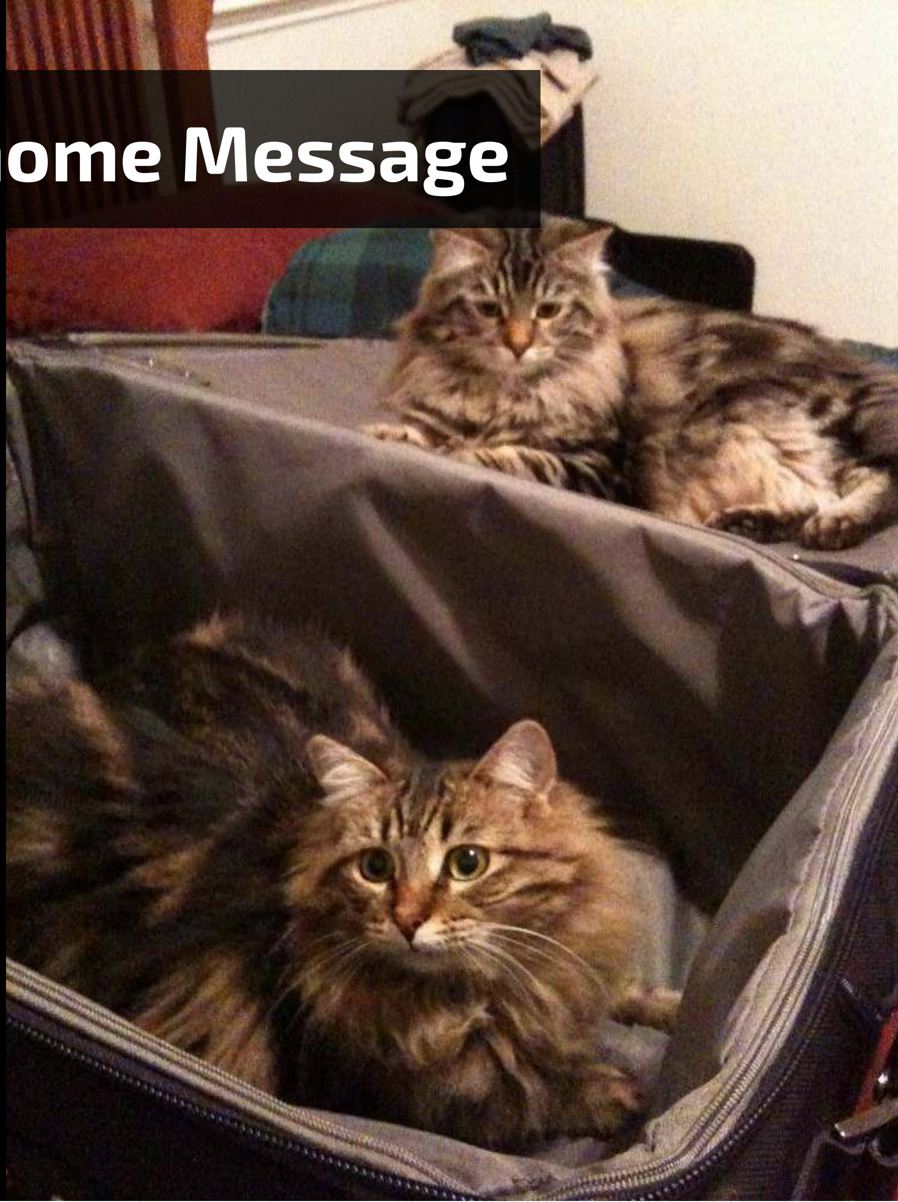




Mixture Models and Alternatives to Model Selection (skipped)

Learning Sparse Matrices (skipped)

Take-home Message



That's It!
Questions?