National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”
Institute of Physics and Technology

Lecture 11

Generative Models for Discrete Data. Sparse Linear Models

Dmytro Progonov,

PhD, Associate Professor,

Department Of Physics and Information Security Systems

Content

- Problem statement;
- Greedy search;
- Basics of ℓ_1 −regularization;
- Algorithms for ℓ_1 −regularization;
- Extensions of ℓ_1 −regularization.

Problem statement

There are some application where feature selection/sparsity is useful:

- To find *the smallest set of features* that can accurately predict the response in order to prevent overfitting, to reduce the cost of building a diagnosis device, or to help with scientific insight into the problem;
- To select a *subset of the training examples*, which can help reduce overfitting and computational cost (***sparse kernel machine***);
- To find a *sparse representation of signals*, in terms of a small number of some predefined basis functions.

Greedy search

Suppose we want to find the MAP model. If we use the ℓ_0 −regularized objective (related to number of non-zero elements of the vector), we can exploit properties of least squares to derive various *efficient greedy forward search methods*, such as:

- Single best replacement;
- Orthogonal least squares;
- Orthogonal matching pursuits;
- Matching pursuits;
- Backward selection;
- Forward-Backwards (FoBa) algorithm;
- Stochastic search;
- EM and variational inference.

Basics of ℓ_1 −regularization (1/1)

Although greedy algorithms often work well, they can of course *get stuck in local optima*. Part of the problem is due to the fact the feature vector's element γ_j are discrete $\gamma_j \in \{0; 1\}$. It is common to *relax hard constraints* of this form by *replacing discrete with continuous variables*.

Consider a prior of the form:

$$p(\mathbf{w}|\lambda) = \prod_{j=1}^{D} Lap\left[w_j|0, 1/\lambda\right] \propto \prod_{j=1}^{D} e^{-\lambda|w_j|}$$

We will use a uniform prior on the offset term, $p(w_0) \propto 1$. Let us perform MAP estimation with this prior. The penalized negative log likelihood has the form

$$f(\mathbf{w}) = -\log[p(\mathcal{D}|\mathbf{w})] - \log[p(\mathbf{w}|\lambda)] = NLL(\mathbf{w}) + \lambda\|\mathbf{w}\|_1$$

where $\|\mathbf{w}\|_1 = \sum_{j=1}^{D}\left|w_j\right|$ is the ℓ_1 norm of $\mathbf{w}$.

Basics of ℓ_1 −regularization (2/2)

For suitable large λ, the estimate $\hat{\mathbf{w}}$ will be sparse. Indeed, this can be thought of as a convex approximation to the non-convex ℓ_0 objective

$$\boxed{\underset{\mathbf{w}}{\operatorname{argmin}}\, NLL(\mathbf{w}) + \lambda\|\mathbf{w}\|_0}$$

In the case of linear regression, the ℓ_1 objective becomes

$$\boxed{f(\mathbf{w}) = \sum_{i=1}^{N}\left(-\frac{1}{2\sigma^2}\right)(y_i - (w_0 + \mathbf{w}^T\mathbf{x}_i))^2 + \lambda\|\mathbf{w}\|_1 = RSS(\mathbf{w}) + \acute{\lambda}\|\mathbf{w}\|_1}$$

where $\acute{\lambda} = 2\lambda\sigma^2$. This method is known as ***<u>basis pursuit denoising</u>*** (***<u>BPDN</u>***).

Comparison of least squares, lasso, ridge and subset selection

We can write down the MAP and ML estimates analytically, as follows:

MLE – the OLS solution is given by

$$\boxed{\hat{w}_k{}^{OLS} = \mathbf{x}_k{}^T\mathbf{y}}$$

Ridge – ridge estimate is given by

$$\boxed{\hat{w}_k{}^{ridge} = \frac{\hat{w}_k{}^{OLS}}{1+\lambda}}$$

Lasso – lasso estimate is given by

$$\boxed{\hat{w}_k{}^{lasso} = sign\left[\hat{w}_k{}^{OLS}\right]\left(\left|\hat{w}_k{}^{OLS}\right| - \frac{\lambda}{2}\right)_+}$$

Subset selection – if we pick the best K features, the parameter estimate is as follow

$$\boxed{\hat{w}_k{}^{SS} = \begin{cases} \hat{w}_k{}^{OLS} & if\ rank\left(\left|\hat{w}_k{}^{OLS}\right|\right) < K \\ 0 & otherwise \end{cases}}$$

Algorithms for ℓ_1 −regularization

Coordinate descent

$$\boxed{{w_j}^* = \underset{z}{\operatorname{argmin}} f(\mathbf{w} + z\mathbf{e}_j) - f(\mathbf{w})}$$

Least angle regression and shrinkage (LARS)

Proximal and gradient projection methods

Proximal gradient method

Nesterov's method

EM for lasso

The E step

$$\boxed{p(1/{\tau_j}^2 \,|\mathbf{w}, \mathcal{D}) = InverseGaussian\left(\sqrt{\frac{\gamma^2}{{w_j}^2}}, \gamma^2\right)}$$

The M step

$$\boxed{\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} \left[-\frac{1}{2}\overline{\mathbf{w}}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2{}^2 - \frac{1}{2}\mathbf{w}^T\overline{\mathbf{\Lambda}}\mathbf{w}\right]}$$

Extensions of ℓ_1 −regularization

1. Group lasso:
   1. Multinomial logistic regression;
   2. Linear regression with categorical inputs;
   3. Multi-task learning;
2. Fused lasso;
3. Elastic net (ridge and lasso combined);
4. Hierarchical adaptive lasso (as part of non-convex regularizers class).

Conclusion

- Greedy search methods for Sparse Linear Models are presented;
- Basics algorithms of ℓ_1 −regularization were considered;
- Extensions of ℓ_1 −regularization was shown.