Linear Regression and Regularization

CSE 4334 / 5334 Data Mining
Spring 2019

Won Hwa Kim

(Slides courtesy of Mark Craven at UW-Madison)

Feature selection via shrinkage (regularization)

- instead of explicitly selecting features, in some approaches we can bias the learning process towards using a small number of features
- key idea: objective function has two parts
  - term representing error minimimization
  - term that “shrinks” parameters toward 0

Linear Regression

- consider the case of linear regression

$$f(\boldsymbol{x}) = w_0 + \sum_{i=1}^{n} x_i w_i$$

- the standard approach minimizes sum squared error

$$\begin{aligned} E(\boldsymbol{w}) &= \sum_{d \in D} \left(y^{(d)} - f\left(\boldsymbol{x}^{(d)}\right) \right)^2 \\ &= \sum_{d \in D} \left(y^{(d)} - w_0 - \sum_{i=1}^{n} x_i^{(d)} w_i \right)^2 \end{aligned}$$

Ridge Regression and LASSO

- Ridge regression adds a penalty term, the L_2 norm of the weights

$$E(\boldsymbol{w}) = \sum_{d \in D} \left(y^{(d)} - w_0 - \sum_{i=1}^{n} x_i^{(d)} w_i \right)^2 + \lambda \sum_{i=1}^{n} w_i^2$$

- the Lasso method adds a penalty term, the L_1 norm of the weights

$$E(\boldsymbol{w}) = \sum_{d \in D} \left(y^{(d)} - w_0 - \sum_{i=1}^{n} x_i^{(d)} w_i \right)^2 + \lambda \sum_{i=1}^{n} |w_i|$$

LASSO Optimization

- LASSO

$$\arg\min_{\boldsymbol{w}} \sum_{d \in D} \left(y^{(d)} - w_0 - \sum_{i=1}^{n} x_i^{(d)} w_i \right)^2 + \lambda \sum_{i=1}^{n} |w_i|$$

- this is equivalent to the following constrained optimization problem (we get the formulation above by applying the method of Lagrange multipliers to the formulation below)

$$\arg\min_{\boldsymbol{w}} \sum_{d \in D} \left(y^{(d)} - w_0 - \sum_{i=1}^{n} x_i^{(d)} w_i \right)^2 \quad \text{subject to} \quad \sum_{i=1}^{n} |w_i| \le t$$

Ridge regression and the LASSO

β's are are the weights in this figure

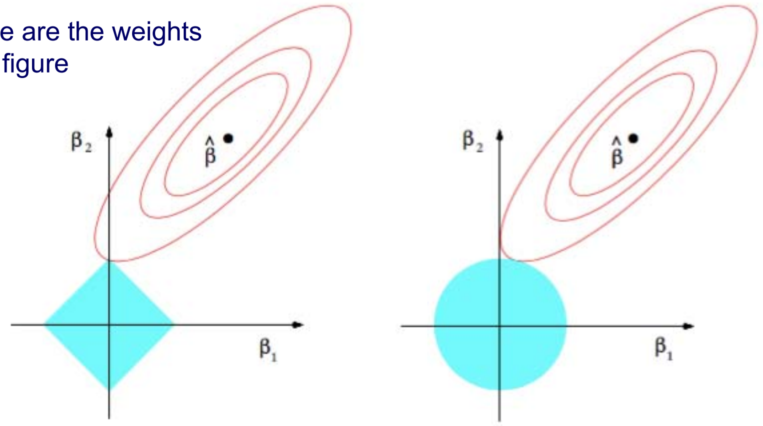


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ and $\beta_1^2 + \beta_2^2 \le t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

Figure from Hastie et al., The Elements of Statistical Learning, 2008

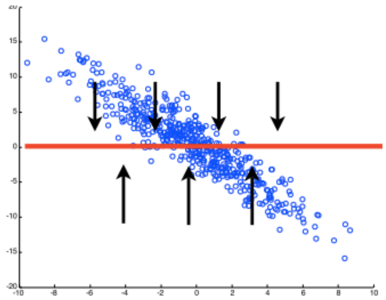
Feature Selection via Shrinkage

- Lasso (L_1) tends to make many weights 0, inherently performing feature selection
- Ridge regression (L_2) shrinks weights but isn't as biased towards selecting features
- L_1 and L_2 penalties can be used with other learning methods (logistic regression, neural nets, SVMs, etc.)
- both can help avoid overfitting by reducing variance
- there are many variants with somewhat different biases
  - elastic net: includes L_1 and L_2 penalties
  - group lasso: bias towards selecting defined groups of features
  - fused lasso: bias towards selecting "adjacent" features in a defined chain
  - etc.

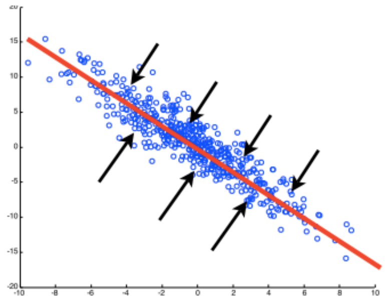
Dimension Reduction

- *feature selection*: equivalent to projecting feature space to a lower dimensional subspace perpendicular to removed feature
- *dimensionality reduction*: allow other kinds of projection (e.g. PCA re-represents data using linear combinations of original features)

feature selection

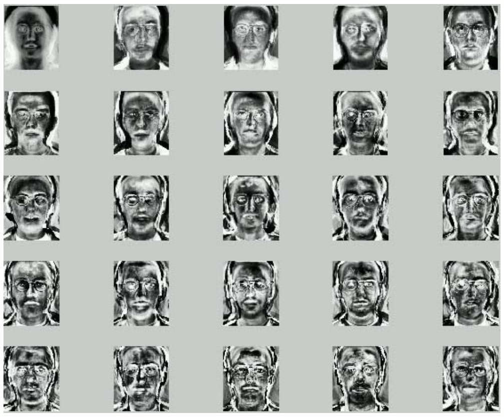


dimensionality reduction



Dimensionality reduction

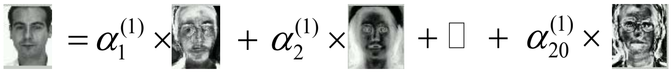
We can represent a face using all of the pixels in a given image (# features = # pixels)



More effective method: represent each face as a linear combination of *eigenfaces* (# features = 20)

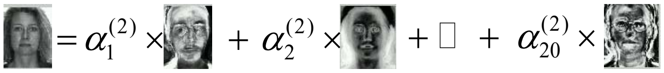
Dimensionality reduction example

represent each face as a linear combination of *eigenfaces*



$$= \alpha_1^{(1)} \times \quad + \alpha_2^{(1)} \times \quad + \square + \alpha_{20}^{(1)} \times$$

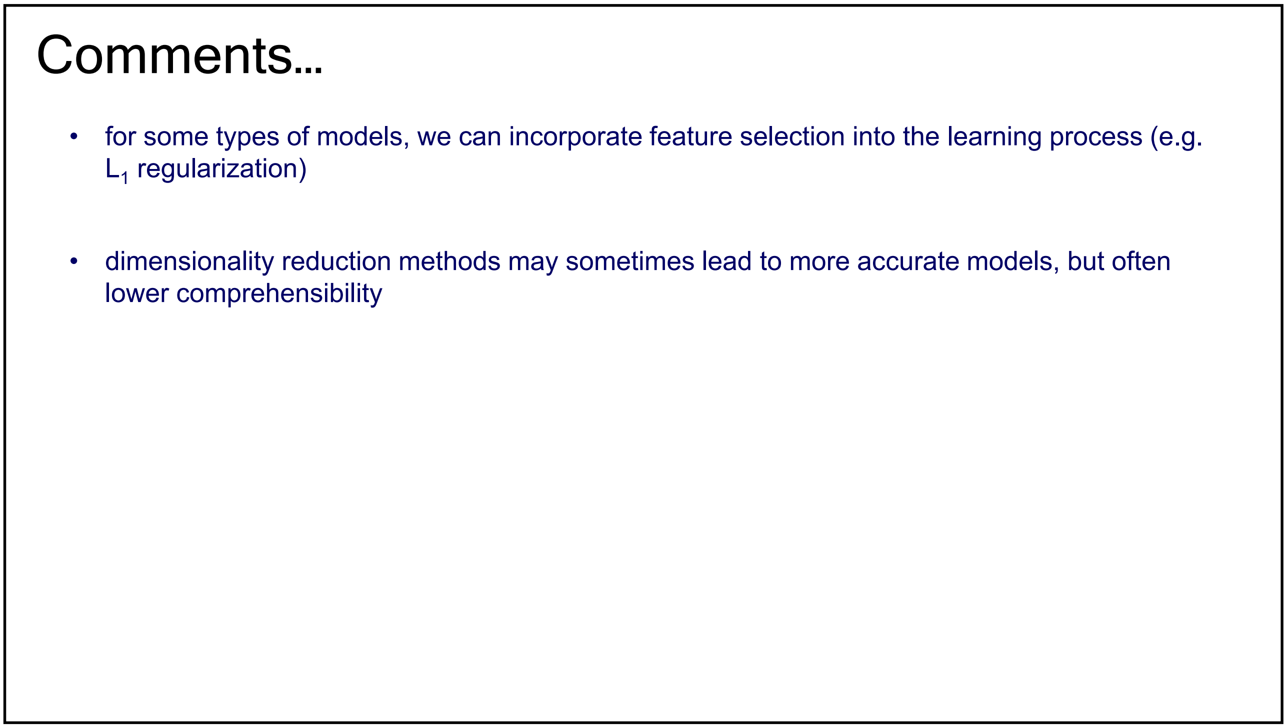
$$\boldsymbol{x}^{(1)} = \left\langle \alpha_1^{(1)},\ \alpha_2^{(1)},\ \square\ ,\ \alpha_{20}^{(1)} \right\rangle$$



$$= \alpha_1^{(2)} \times \quad + \alpha_2^{(2)} \times \quad + \square + \alpha_{20}^{(2)} \times$$

$$\boldsymbol{x}^{(2)} = \left\langle \alpha_1^{(2)},\ \alpha_2^{(2)},\ \square\ ,\ \alpha_{20}^{(2)} \right\rangle$$

# of features is now 20 instead of # of pixels in images



Comments…

- for some types of models, we can incorporate feature selection into the learning process (e.g. L_1 regularization)

- dimensionality reduction methods may sometimes lead to more accurate models, but often lower comprehensibility