Machine Learning for Biomedical Engineering

Enrico Grisan
enrico.grisan@dei.unipd.it
Curse of dimensionality

• Why are more features bad?

  – Redundant features (useless or confounding)
  – Hard to interpret and visualize
  – Hard to store and process
  – Complexity of decision rules (boundaries) increases with the number of features
Occam’s Razor

William of Ockham (1285-1349)
Principle of Parsimony

«One should not increase, beyond what it is necessary, the number of entities required to explain anything»

Seek the simplest solution!
Dimensionality reduction

- Features selection
- Latent features
Best subset selection

- $D$ features
- Find the subset of dimension $k = \{1, \ldots, D\}$ with smallest residual sum
- Exhaustive search:
  \[
  \sum_{k=1}^{D} \binom{D}{k} = \sum_{k=1}^{D} \frac{D!}{k! (D-k)!}
  \]
  There exist methods efficient only for $D < 40$
One-step feature selection

Estimate training or cross-validation accuracy for each single-feature classifier:

\[ \hat{f}_k : X_k \rightarrow y \]
\[ e_k = E[\hat{f}_k(X_k) - y] \]

Choose the set of \( K \) features with smaller error \( e_k \)
Forward stepwise selection

Incrementally add one feature at the time to the current selected set

Greedy algorithm

1. Given a subset $S$ of $K$ features
2. Evaluate the RSS of the subset $S$ augmented with one feature (among the remaining $D - K$)
3. Add the feature providing the larger reduction in RSS
4. Repeat 1-3
Backward stepwise selection

Incrementally remove one feature at the time to the current selected set

1. Start with the regression/classifier estimate on the full set of D features
2. Evaluate the feature providing the less impact (z-score or RSS with one feature removed)
3. Remove the feature from the set of active features
4. Repeat steps 2-3
Forward stagewise

Incrementally add the feature most correlated with the residuals

1. Center and standardize all features
2. Start with intercept equal to \( \bar{y} \)

3. Find the feature \( X_k \) most correlated with the residuals \( r \)
4. Compute the weight \( w'_k \) of the linear regression of \( X_k \) to \( r \)
5. Update \( w_k = w_k + w'_k \)
6. Update the least squares fit and the residuals \( r = \hat{y} - y \)
7. Repeat step 3-5 until no feature is correlated with the residual
Least angle regression (LAR)

1. Center and standardize all features
2. Start with $r = \hat{y} - \bar{y}$ and $w_i = 0$, $i = 1, ..., D$

3. Find the feature $X_k$ most correlated with $r$ and add it to the active set $A^0$

4. At each iteration $\tau$ evaluate the least squares direction:
   \[ \delta_\tau = (X_{A^\tau}^T X_{A^\tau})^{-1} X_{A^\tau}^T r^\tau \]

5. Update the weights of the features in the active set
   \[ w_{A^{\tau+1}} = w_{A^{\tau}} + \eta \delta_\tau \]

6. Evaluate the least squares fit of $X_{A^\tau}$ and update the residuals $r^\tau$

7. Repeat 4-6 until some other variable $X_j$ is correlated to $r^\tau$ as much as $X_{A^\tau}$

8. Add $X_j$ to $X_{A^\tau}$

9. Repeat 4-8
Incremental Forward Stagewise regression

1. Center and standardize all features
2. Start with $r = y$ and $w_i = 0$, $i = 1, ..., D$

3. Find the feature $X_k$ most correlated with $r$
4. Evaluate the change
   \[ \delta = \varepsilon \cdot \text{sign}(\langle X_k, r \rangle)_T \]
5. Update the weights of the features in the active set
   \[ w_k = w_k + \delta \]
6. Update the residuals
   \[ r = r - \delta X_k \]
7. Repeat 3-6 until the residuals are uncorrelated with the features
Comparison

![Comparison Graph]

- Forward Stepwise
- LAR
- Lasso
- Forward Stagewise
- Incremental Forward Stagewise

Fraction of $L_1$ arc-length

$E ||\hat{\beta}(k) - \beta||^2$
fMRI experiment

Test whether a classifier could distinguish the activation as a result of seeing words that were either kinds of tool or kinds of building. The subject was shown one word per trial and performed the following task: the subject should think about the item and its properties while the word was displayed (3 s) and try to clear her mind afterwards (8s of blank screen).

(Pereira, Mitchell, Botvinick, Neuroimage, 2009)

For each patient and each task (recognizing word) fMRI data provide a signal correlated to metabolism at each voxel of the acquired 3D barin volume

16000 features/ example
42 example s for the «building» class
42 examples for the «tools» class
fMRI feature selection

The goal is:
- reduce the ratio of features to examples,
- decrease the chance of overfitting,
- get rid of uninformative features
- let the classifier focus on informative ones.
# fMRI feature selection

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of voxels</th>
<th>NCV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>Searchlight</td>
<td>0.81</td>
<td>0.82</td>
</tr>
<tr>
<td>Activity</td>
<td>0.79</td>
<td>0.80</td>
</tr>
<tr>
<td>ANOVA</td>
<td>0.77</td>
<td>0.75</td>
</tr>
</tbody>
</table>

16000 features!!!
Regularization

- Do not need validation set to know some fits are silly

- Discourage solutions we don’t like

- Formalizing the cost of solutions we do not like
Shrinkage

Minimizing the objective function

$$\hat{w} = \arg\min_w L(\hat{y}(x_*); y_*) + \lambda \|w\|_p$$

Or constraining the estimates

$$\begin{cases} \hat{w} = \arg\min_w L(\hat{y}(x_*); y_*) \\ \|\hat{w}\|_p = t \end{cases}$$
Ridge regression

\[ \hat{w}_{ridge} = \arg \min_w L(\hat{y}(x_*); y_*) + \lambda ||w||_2 \]

\[ \hat{w}_{ridge} = \arg \min_w \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{D} w_j^2 \]

\[ RSS(\lambda) = (y - Xw)^T (y - Xw) + \lambda w^T w \]

\[ \hat{w}_{ridge} = (X^T X + \lambda I)^{-1} X^T y \]
LASSO

Least Absolute Shrinkage and Selection Operator

\[ \hat{w}^{lasso} = \arg \min_w L(\hat{y}(x_*); y_*) + \lambda \|w\|_1 \]

\[ \hat{w}^{lasso} = \arg \min_w \sum_{i=1}^N (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^D |w_j| \]
Geometry of shrinkage

\[ L(\hat{y}(x_*); y_*) \]

\[ \lambda \| w \|_2 \]

\[ \lambda \| w \|_1 \]
Other shrinkage norms

\[ p = 4 \quad p = 2 \quad p = 1 \quad p = 0.5 \quad p = 0.2 \]

\[ p = 1.2 \quad \alpha = 0.2 \]

Elastic net:
\[ \alpha \| \mathbf{w} \|_2 + (1 - \alpha) \| \mathbf{w} \|_1 \]
Regularization constants

How do we pick $\lambda$ or $t$?

1) Based on validation
2) Based on bounds on generalization error
3) Based on empirical Bayes
4) Reinterpreting $\lambda$
5) Going full Bayesian approach
Pre processing

Centering
- Might have all features at 500 ± 10
- Hard to predict ball park of bias
- Subtract mean from all input features

Rescaling
- Heights can be measured in cm or m
- Rescale inputs to have unit variance
  ... or interquartile ranges

Care at test time:
apply same scale to all inputs and reverse scaling to prediction
Some tricks of the trade

• Preprocessing

• Transformations

• Features
Log transform inputs

Positive quantities are often highly skewed
Log-domain id often much more natural
Creating extra data

Dirty trick:
Create more training ‘data’ by corrupting examples in the real training set

Changes could respect invariances that would be difficult or burdensome to measure directly
Encoding attributes

• **Categorical variables**
  – A study has three individuals
  – Three different colours
  – Possible encoding: 100, 010, 001

• **Ordinal variables**
  – Movie rating, stars
  – Tissue anomaly rating, expert scores 1-3
  – Possible encoding: 00, 10, 11
Basis function features

In the regression and classification examples we used polynomials \( x, x^2, x^3, ... \)

Often a bad choice

Polynomials of sparse binary features may make sense:
\( x_1 x_2, x_1 x_3, ..., x_1 x_2 x_2 \)

Other options:
- radial basis function: \( e^{-|x-\mu|^2/h^2} \)
- sigmoids: \( 1/(1 + e^{-v^T x}) \)
- Fouries, wavelets ...