15-388/688 - Practical Data Science: Nonlinear modeling, cross-validation, and regularization

J. Zico Kolter
Carnegie Mellon University
Spring 2018
Outline

Example: return to peak demand prediction

Overfitting, generalization, and cross validation

Regularization

General nonlinear features

Kernels

Nonlinear classification
Announcements

Tutorial “proposal” sentence due tonight

I will send feedback on topics by next week, you *may* change topics after feedback, but don’t submit with the intention of doing this

Piazza note about linear regression in HW 3

TA Office Hours calendar posted to course webpage, under “Instructors”
Example: return to peak demand prediction

Overfitting, generalization, and cross validation

Regularization

General nonlinear features

Kernels

Nonlinear classification
Peak demand vs. temperature (summer months)
Peak demand vs. temperature (all months)
Linear regression fit

- Observed days
- Squared loss fit

Graph showing relationship between Peak Demand (GW) and High Temperature (F).
“Non-linear” regression

Thus far, we have illustrated linear regression as “drawing a line through the data”, but this was really a function of our input features.

Though it may seem limited, linear regression algorithms are quite powerful when applied to non-linear features of the input data, e.g.

\[ x^{(i)} = \begin{bmatrix} \text{High\_Temperature}^{(i)} \\ \text{High\_Temperature}^{(i)} \\ 1 \end{bmatrix} \]

Same hypothesis class as before \( h_\theta(x) = \theta^T x \), but now prediction will be a non-linear function of base input (e.g. a quadratic function).

Same least-squares solution \( \theta = (X^T X)^{-1} X^T y \)
Polynomial features of degree 2
The only element we need to add to write this non-linear regression is the creation of the non-linear features.

```python
x = df_daily.loc[:, "Temperature"]
min_x, rng_x = (np.min(x), np.max(x) - np.min(x))
x = 2*(x - min_x)/rng_x - 1.0
y = df_daily.loc[:, "Load"]

X = np.vstack([x**i for i in range(poly_degree,-1,-1)]).T
theta = np.linalg.solve(X.T.dot(X), X.T.dot(y))

x0 = 2*(np.linspace(xlim[0], xlim[1],1000) - min_x)/rng_x - 1.0
X0 = np.vstack([x0**i for i in range(poly_degree,-1,-1)]).T
y0 = X0.dot(theta)
```

Output learned function:
Polynomial features of degree 3

Observed days
Squared loss fit

Peak Demand (GW)

High Temperature (F)
Polynomial features of degree 4
Polynomial features of degree 10

![Graph showing observed days and squared loss fit with high temperature (F) on the x-axis and peak demand (GW) on the y-axis. The graph visually represents the relationship between high temperature and peak demand, with a polynomial fit curve superimposed on the scatter plot of observed data points.]
Polynomial features of degree 50
Linear regression with many features

Suppose we have $m$ examples in our data set and $n = m$ features (plus assumption that features are linearly independent, though we’ll always assume this)

Then $X \in \mathbb{R}^{m \times n}$ is a square matrix, and least squares solution is:

$$
\theta = (X^T X)^{-1} X^T Y = X^{-1} X^{-T} X^T y = X^{-1} y
$$

and we therefore have $X\theta = y$ (i.e., we fit data exactly)

Note that we can only perform the above operations when $X$ is square, though if we have more features than examples, we can still get an exact fit by simply discarding features
Outline

Example: return to peak demand prediction

Overfitting, generalization, and cross validation

Regularization

General nonlinear features

Kernels

Nonlinear classification
Generalization error

The problem we the canonical machine learning problem is that we don’t really care about minimizing this objective on the given data set

\[
\min_{\theta} \sum_{i=1}^{m} \ell(h_{\theta}(x^{(i)}), y^{(i)})
\]

What we really care about is how well our function will generalize to new examples that we didn’t use to train the system (but which are drawn from the “same distribution” as the examples we used for training)

The higher degree polynomials exhibited overfitting: they actually have very low loss on the training data, but create functions we don’t expect to generalize well
Cartoon version of overfitting

As model becomes more complex, training loss always decreases; generalization loss decreases to a point, then starts to increase.
Cross-validation

Although it is difficult to quantify the true generalization error (i.e., the error of these algorithms over the complete distribution of possible examples), we can approximate it by **holdout cross-validation**

Basic idea is to split the data set into a training set and a holdout set

<table>
<thead>
<tr>
<th>Training set (e.g. 70%)</th>
<th>Holdout / validation set (e.g. 30%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All data</td>
<td></td>
</tr>
</tbody>
</table>

Train the algorithm on the training set and evaluate on the holdout set
Cross-validation in code

A simple example of holdout cross-validation:

```python
# compute a random split of the data
np.random.seed(0)
perm = np.random.permutation(len(df_daily))
idx_train = perm[:int(len(perm)*0.7)]
idx_cv = perm[int(len(perm)*0.7):]

# scale features for each split based upon training
xt = df_daily.iloc[idx_train, 0]
min_xt, rng_xt = (np.min(xt), np.max(xt) - np.min(xt))
xt = 2*(xt - min_xt)/rng_xt - 1.0
xcv = 2*(df_daily.iloc[idx_cv, 0] - min_xt)/rng_xt - 1
yt = df_daily.iloc[idx_train, 1]
ycv = df_daily.iloc[idx_cv, 1]

# compute least squares solution and error on holdout and training
X = np.vstack([xt**i for i in range(poly_degree,-1,-1)]).T
theta = np.linalg.solve(X.T.dot(X), X.T.dot(yt))
err_train = 0.5*np.linalg.norm(X.dot(theta) - yt)**2/len(idx_train)
err_cv = 0.5*np.linalg.norm(Xcv.dot(theta) - ycv)**2/len(idx_cv)
```
Parameters and hyperparameters

We refer to the $\theta$ variables as the parameters of the machine learning algorithm.

But there are other quantities that also affect the classifier: degree of polynomial, amount of regularization, etc; these are collectively referred to as the hyperparameters of the algorithm.

Basic idea of cross-validation: use training set to determine the parameters, use holdout set to determine the hyperparameters.
Illustrating cross-validation
Training and cross-validation loss by degree
Training and cross-validation loss by degree

![Graph showing training and cross-validation loss by degree]

- **Y-axis:** Average Squared Loss
- **X-axis:** Polynomial Degree
- **Legend:**
  - Blue: Training set
  - Green: Holdout set
K-fold cross-validation

A more involved (but actually slightly more common) version of cross validation

Split data set into $k$ disjoint subsets (folds); train on $k - 1$ and evaluate on remaining fold; repeat $k$ times, holding out each fold once

Report average error over all held out folds
Variants

**Leave-one-out cross-validation:** the limit of k-fold cross-validation, where each fold is only a single example (so we are training on all other examples, testing on that one example)

[Somewhat surprisingly, for least squares this can be computed *more* efficiently than k-fold cross validation, same complexity solving for the optimal $\theta$ using matrix equation]

**Stratified cross-validation:** keep an approximately equal percentage of positive/negative examples (or any other feature), in each fold

**Warning:** k-fold cross validation is *not* always better (e.g., in time series prediction, you would want to have holdout set all occur after training set)
Outline

Example: return to peak demand prediction

Overfitting, generalization, and cross validation

Regularization

General nonlinear features

Kernels

Nonlinear classification
Regularization

We have seen that the degree of the polynomial acts as a natural measure of the “complexity” of the model, higher degree polynomials are more complex (taken to the limit, we fit any finite data set exactly)

But fitting these models also requires extremely large coefficients on these polynomials

For 50 degree polynomial, the first few coefficients are
\[ \theta = -3.88 \times 10^6, 7.60 \times 10^6, 3.94 \times 10^6, -2.60 \times 10^7, \ldots \]

This suggests an alternative way to control model complexity: keep the weights small (regularization)
This leads us back to the regularized loss minimization problem we saw before, but with a bit more context now:

\[
\text{minimize } \sum_{i=1}^{m} \ell(h_{\theta}(x^{(i)}), y^{(i)}) + \frac{\lambda}{2} \|\theta\|_2^2
\]

This formulation trades off loss on the \textit{training} set with a penalty on high values of the parameters.

By varying \(\lambda\) from zero (no regularization) to infinity (infinite regularization, meaning parameters will all be zero), we can sweep out different sets of model complexity.
Regularized least squares

For least squares, there is a simple solution to the regularized loss minimization problem

$$\text{minimize}_\theta \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})^2 + \lambda \|\theta\|_2^2$$

Taking gradients by the same rules as before gives:

$$\nabla_\theta \left( \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})^2 + \lambda \|\theta\|_2^2 \right) = 2X^T (X\theta - y) + 2\lambda \theta$$

Setting gradient equal to zero leads to the solution

$$2X^T X\theta + 2\lambda \theta = 2X^T y \implies \theta = (X^T X + \lambda I)^{-1} X^T y$$

Looks just like the normal equations but with an additional $\lambda I$ term
50 degree polynomial fit

Observed days
Squared loss fit
50 degree polynomial fit – $\lambda = 1$
Training/cross-validation loss by regularization
Training/cross-validation loss by regularization

![Graph showing training and holdout set losses vs regularization parameter.]
Suppose you run linear regression with polynomial features and some initial guess for $d$ and $\lambda$. You find that your validation loss is much higher than your training loss. Which actions might be beneficial to take?

1. Decrease $\lambda$
2. Increase $\lambda$
3. Decrease $d$
4. Increase $d$
Outline

Example: return to peak demand prediction

Overfitting, generalization, and cross validation

Regularization

General nonlinear features

Kernels

Nonlinear classification
Notation for more general features

We previously described polynomial features for a *single* raw input, but if our raw input is itself multi-variate, how do we define polynomial features?

Deviating a bit from past notion, for precision here we’re going to use $x^{(i)} \in \mathbb{R}^k$ to denote the *raw* inputs, and $\phi^{(i)} \in \mathbb{R}^n$ to denote the input features we construct (also common to use the notation $\phi(x^{(i)})$)

We’ll also drop $(i)$ superscripts, but important to understand we’re transforming each feature this way

E.g., for the high temperature:

$$x = [\text{High\_Temperature}], \quad \phi = \begin{bmatrix} x^2 \\ x \\ 1 \end{bmatrix}$$
Polynomial features in general

One possibility for higher degree polynomials is to just use an independent polynomial over each dimension (here of degree $d$)

$$x \in \mathbb{R}^k \implies \phi = \begin{bmatrix} x_1^d \\ \vdots \\ x_1 \\ \vdots \\ x_k^d \\ \vdots \\ x_k \\ 1 \end{bmatrix} \in \mathbb{R}^{kd+1}$$

But this ignores cross terms between different features, i.e., terms like $x_1 x_2^2 x_k$
Polynomial features in general

A better generalization of polynomials is to include all polynomial terms between raw inputs up to degree $d$

$$x \in \mathbb{R}^k \implies \phi = \left\{ \prod_{i=1}^{k} x_i^{b_i} : \sum_{i=1}^{n} b_i \leq d \right\} \in \mathbb{R}^{\binom{k+d}{k}}$$

Code to generate all polynomial features with degree exactly $d$:

```python
from itertools import combinations_with_replacement

[np.prod(a) for a in combinations_with_replacement(x, d)]
```

Code to generate all polynomial features with degree up to $d$

```python
[np.prod(a) for i in range(d+1) for a in combinations_with_replacement(x, i)]
```
Code for general polynomials

The following code efficiently (relatively) generates all polynomials up to degree $d$ for an entire data matrix $X$

```python
def poly(X,d):
    return np.array([reduce(operator.mul, a, np.ones(X.shape[0])))
                     for i in range(1,d+1)  
                     for a in combinations_with_replacement(X.T, i)]).T
```

It is using the same logic as above, but applying it to entire columns of the data at a time, and thus only needs one call to combinations_with_replacement
Radial basis functions (RBFs)

For $x \in \mathbb{R}^k$, select some set of $p$ centers, $\mu^{(1)}, \ldots, \mu^{(p)}$ (we’ll discuss shortly how to select these), and create features

$$\phi = \left\{ \exp \left( -\frac{\|x - \mu^{(i)}\|^2}{2\sigma^2} \right) : i = 1, \ldots, p \right\} \cup \{1\} \in \mathbb{R}^{p+1}$$

**Very important:** need to normalize columns of $X$ (i.e., different features), to all be the same range, or distances won’t be meaningful.

(Hyper)parameters of the features include the choice of the $p$ centers, and the choice of the *bandwidth* $\sigma$.

Choose centers, i.e., to be a uniform grid over input space, can choose $\sigma$ e.g. using cross validation (don’t do this, though, more on this shortly).
Example radial basis function

Example:

\[ x = [\text{High\_Temperature}], \]

\[ \mu^{(1)} = [20], \mu^{(2)} = [25], \ldots, \mu^{(16)} = [95], \sigma = 10 \]

Leads to features:

\[ \phi = \begin{bmatrix}
\exp(-(\text{High\_Temperature} - 20)^2/200) \\
\vdots \\
\exp(-(\text{High\_Temperature} - 95)^2/200) \\
1
\end{bmatrix} \]
The following code generates a complete set of RBF features for an entire data matrix $X \in \mathbb{R}^{m \times k}$ and matrix of centers $\mu \in \mathbb{R}^{p \times k}$.

```python
def rbf(X, mu, sig):
    sqdist = -2*X@mu.T + (X**2).sum(axis=1)[:, None] + (mu**2).sum(axis=1)
    return np.exp(-sqdist/(2*sig**2))
```

Important “trick” is to efficiently compute distances between all data points and all centers.
Poll: complexity of computing features

For \( n \) dimensional input, \( m \) examples, \( p \) centers, what is the complexity of computing the RBF feature \( \phi \) for every training example \( x^{(i)} \)?

1. \( O(mnp) \)

2. \( O(mp) \)

3. \( O(mn^2p) \)

4. \( O(mn^2p^2) \)
Difficulties with general features

The challenge with these general non-linear features is that the number of potential features grows very quickly in the dimensionality of the raw input.

**Polynomials:** $k$-dimensional raw input $\implies \binom{k + d}{k} = O(d^k)$ total features (for fixed $d$)

**RBFs:** $k$-dimensional raw input, uniform grid with $d$ centers over each dimension $\implies d^k$ total features

These quickly become impractical for large feature raw input spaces.
Practical polynomials

Don't use the full set of all polynomials, for anything but very low dimensional input data (say $k \leq 4$)

Instead, form polynomials only of features where you know that the relationship may be important:

E.g. $\text{Temperature}^2 \cdot \text{Weekday}$, but not $\text{Temperature} \cdot \text{Humidity}$

For binary raw inputs, no point in taking every power ($x_i^2 = x_i$)

These elements do all require some insight into the problem
Practical RBFs

Don’t create RBF centers in a grid over your raw input space (your data will never cover an entire high-dimensional space, but will lie on a subset).

Instead, pick centers by randomly choosing $p$ data points in the training set (a bit fancier, run k-means to find centers, which we’ll describe later).

Don’t pick $\sigma$ using cross validation.

Instead, choose the following (called the median trick):

$$\sigma = \text{median}(\{\|\mu^{(i)} - \mu^{(j)}\|_2, i, j = 1, ..., p\})$$
Example: return to peak demand prediction

Overfitting, generalization, and cross validation

Regularization

General nonlinear features

Kernels

Nonlinear classification
Kernels

One of the most prominent advances in machine learning in the past 20 years (recently fallen out of favor relative to neural networks, but still can be the best-performing approach for many “medium-sized” problems)

Kernels fundamentally are about specific hypothesis function

\[ h_\theta(x) = \sum_{i=1}^{m} \theta_i K(x, x^{(i)}) \]

where \( K : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R} \) is a kernel function

Kernels can implicitly represent high dimensional feature vectors without the need to form them explicitly (we won’t prove this here, but will provide a short description in the notes over break)
Kernels as high dimensional features

1. Polynomial Kernel

\[ K(x, z) = (1 + x^T z)^d \]

is equivalent to using full degree \( d \) polynomial \( \binom{n+d}{d} \)-dimension features in the raw inputs

2. RBF Kernel

\[ K(x, z) = \exp\left(-\frac{\|x - z\|_2^2}{2\sigma^2}\right) \]

is equivalent to an *infinite dimensional* RBF feature with centers at every point in space
Kernels: what is the “catch”

What is the downside of using kernels?

Recall hypothesis function

\[ h_\theta(x) = \sum_{i=1}^{m} \theta_i K(x, x^{(i)}) \]

Note that we need a parameter for every training example (complexity increases with the size of the training set)

Called a non-parametric method (number of parameters increase with the number of data points)

Typically, complexity of resulting ML algorithm is \( O(m^2) \) (or larger), leads to impractical algorithms on large data sets
Poll: complexity of gradient descent with kernels

Given by kernel hypothesis function

\[ h_\theta(x) = \sum_{i=1}^{m} \theta_i K(x, x^{(i)}) \]

and the RBF kernel, what is the complexity of computing the gradient of the machine learning objective

\[ \nabla_\theta \sum_{i=1}^{m} \ell(h_\theta(x^{(i)}), y^{(i)}) \]?

1. \( O(mn) \)
2. \( O(mn^2) \)
3. \( O(m^2n) \)
4. \( O(m^2n^2) \)
Outline

Example: return to peak demand prediction

Overfitting, generalization, and cross validation

Regularization

General nonlinear features

Kernels

Nonlinear classification
Nonlinear classification

Just like linear regression, the nice thing about using nonlinear features for classification is that our algorithms remain exactly the same as before. I.e., for an SVM, we just solve (using gradient descent)

$$\min_{\theta} \sum_{i=1}^{m} \max \{ 1 - y^{(i)} \cdot \theta^T x^{(i)}, 0 \} + \frac{\lambda}{2} \| \theta \|^2$$

Only difference is that $x^{(i)}$ now contains non-linear functions of the input data.
Linear SVM on cancer data set

![Graph showing Linear SVM on cancer data set with separating boundary and data points labeled as Benign and Malignant.](image-url)
Polynomial features $d = 2$
Polynomial features $d = 3$
Polynomial features $d = 10$
Below, we assume that $X$ has been normalized so that each feature lies between $[-1, +1]$ (same as we did for polynomial features)

We’re consider to observe how the classifier changes as we change different parameters of the RBFs

$p$ will refer to total number of centers, $d$ will refer the number of centers along each dimensions, assuming centers form a regular grid (so since we have two raw inputs, $p = d^2$)
RBF features, $d = 3, \sigma = 2/d$
RBF features, $d = 10, \sigma = \frac{2}{d}$
RBF features, $d = 20, \sigma = 2/d$
Model complexity and bandwidth

We can control model complexity with RBFs in three ways: two of which we have already seen

1. Choose number of RBF centers
2. Increase/decrease regularization parameter
3. Increase/decrease bandwidth
RBF features, $d = 20$, $\sigma = 0.1$
RBF features, $d = 20$, $\sigma = 0.5$
RBF features, $d = 20, \sigma = 1.07$ (median trick)
RBFs from data, $p = 50, \sigma = \text{median_trick}$