Machine learning notes

Richard Bene

Contents

1 Basic numerical methods **3**
1.1 Gradient descent . . . 3
1.2 Conjugate gradient descent . . . 3
1.3 Stochastic descent . . . 3
1.4 Normal equation method . . . 3

2 Notation **3**

3 Logistic regression **3**

4 Linear regression **3**
4.1 The problem . . . 3

5 Support vector machine **3**

6 Regularization **3**
6.1 Motivation . . . 3
6.2 Regularization . . . 4
6.3 Optimization objective . . . 4
6.4 More generally . . . 4
6.5 Regularized linear regression . . . 4
6.6 Normal equation method with regularization . . . 4
6.7 Regularized logistic regression . . . 5
6.8 Regularized SVM . . . 5

7 Machine learning algorithm types **5**

8 Active learning **5**

9 Kernels **5**
9.1 Non-linear predictions, motivation for kernels . . . 5
9.2 Kernel form of linear regression . . . 5
9.3 Kernel form of SVM . . . 6
9.4 Examples of kernels . . . 6
9.4.1 Simple kernels . . . 6
9.4.2 Composite kernels . . . 6

9.4.3 String kernels 6
9.5 Kernel optimization 6

10 Combining, boosting classifiers 6

11 Clustering, spectral clustering 6

12 Hidden markov models 6

13 Literature 6

1 Basic numerical methods

to train a model we need these methods, the different approaches will lead to minimization problems, which can be solved by the following methods

1.1 Gradient descent

$J(\theta; \mathbf{X})$ is the error function for a given input set $\mathbf{X}$

$$\theta_{j+1} = \theta_j - \alpha \partial_{\theta_j} J(\theta)$$

1.2 Conjugate gradient descent

1.3 Stochastic descent

1.4 Normal equation method

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

2 Notation

$m =$ number of training samples
$n =$ dimension of the parameter, number of features

3 Logistic regression

4 Linear regression

4.1 The problem

$x^{(i)}, y^{(i)}$ is the i-th training example
our model will be linear:

$$h_\theta(x) = \theta^T x$$

inhomogeous case:

$$h_\theta(x) = \theta^T x + \theta_0$$

5 Support vector machine

6 Regularization

6.1 Motivation

a solution to overfitting
overfitting occurs, when the algorithm doesn't generalize well
eg. 5 points given and we are to model it with a 4th degree polynomial, we

have **too many features**, in this case for a new example we can have poor prediction
underfitting occurs, when we use too less features

6.2 Regularization

it works well if we have a lot of features and each has little effect on the output

6.3 Optimization objective

we penalize some parameters
let M_k be big numbers
the error function is the following, for some $i_0, ..., i_N$

$$J(\theta) = \min_\theta \frac{1}{2m} \left\| h_\theta(x^{(i)}) - y^{(i)} \right\|^2 + \mathbf{M_0}\theta_{\mathbf{i_0}} + ... + \mathbf{M_N}\theta_{\mathbf{i_n}}$$

this way the parameter θ_{i_k} have to be small at the minimum

6.4 More generally

$$J(\theta) = \min_\theta \frac{1}{2m} \Big(\left\| h_\theta(x^{(i)}) - y^{(i)} \right\|^2 + \lambda \left\| \theta \right\|^2 \Big)$$

don't penalize θ_0 constant term
it will keep the parameters small
λ - penalty
if λ is very large $\Rightarrow \theta_0 \neq 0$, but $\theta_k = 0$ for $k > 0 \Rightarrow$ underfitting
m is large $\Rightarrow \frac{\lambda}{m} \to 0 \Rightarrow$ the regularization effect decreases
but we can leave out $\lambda \left\| \theta \right\|^2$ from the parantheses, so it won't scale with $m \Rightarrow$ reguralization effect won't decrease as $m \to \infty$
different penalties: choose different norms l_2, l_1

6.5 Regularized linear regression

$$\begin{aligned} \theta_{j+1} &= \theta_j - \alpha \Big[\frac{1}{m} \sum_{i=0}^{n} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} - \frac{\lambda}{m} \theta_j \Big] \\ &= \theta_j (1 - \alpha \frac{\lambda}{m}) - \frac{\alpha}{m} \sum_{i=0}^{n} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} \end{aligned}$$

the factor $(1 - \alpha \frac{\lambda}{m}) < 1$ is the number which shrinks θ

6.6 Normal equation method with regularization

$m \leq n \Rightarrow X^T X$ is not invertable, but $X^T X + \lambda diag(n+1)$ always invertable

6.7 Regularized logistic regression

6.8 Regularized SVM

7 Machine learning algorithm types

- **discriminative** methods
  separate the whole data by a curve
  learn $p(y|x)$
  eg.: previous methods
- **generative** methods
  for each group describe the group with a model
  for a new example the method check which group model fits better
  learn $p(x|y)$

8 Active learning

choosing the samples that are best for learning from them

9 Kernels

9.1 Non-linear predictions, motivation for kernels

$\epsilon \sim N(0, \sigma^2)$

$$y = \theta^T \phi(x) + \theta_0 + \epsilon$$

with

$$x = [x_1, x_2]^T \rightarrow \phi(x) = [1, x_1, x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2]^T$$

with the dimension of x increasing the dimension of $\phi(x)$ radically increase if we use higher order polynomial expansion. But let's have an another example for $\phi(x)$

$$\phi(x) = [1, \sqrt{3}x, \sqrt{3}x^2, x^3]^T$$
$$\phi(x)^T\phi(x') = 1 + 3xx' + 3(xx')^2 + (xx')^3 = (1 + xx')^3$$

9.2 Kernel form of linear regression

The goal is turn the problem into a form, that involve only inner products between feature vectors.

$$y = \theta^T \phi(x) + \epsilon,$$

where ϕ is a feature expansion. The regularized linear least squares objective to minimize is

$$J(\theta) = \left\| y_t - \theta^T \phi(x_t) \right\|^2 + \lambda \left\| \theta \right\|^2 .$$

The optimal θ can be calculated by setting $\partial_\theta J(\theta) = 0$, this gives

$$\theta = \frac{1}{\lambda} \sum_t \alpha_t \phi(x_t).$$

$$\alpha_{t'} = y_t - \theta^T \phi(x_t) = y_t - \frac{1}{\lambda} \sum_{t'} \alpha_{t'} \phi(x_{t'})^T \phi(x_t)$$

α_t depends only on the actual responses y_t and the inner products between the training examples. *Gram matrix:*

$$K_{ij} = \phi(x_i)^T \phi(x_j)$$

$$a = [\alpha_1, ..., \alpha_n]$$

$$a = y - \frac{1}{\lambda} K a$$

the solution is

$$\hat{a} = \lambda(\lambda I + K)^{-1} y.$$

9.3 Kernel form of SVM

9.4 Examples of kernels

9.4.1 Simple kernels

9.4.2 Composite kernels

9.4.3 String kernels

9.5 Kernel optimization

10 Combining, boosting classifiers

11 Clustering, spectral clustering

12 Hidden markov models

13 Literature

Andrew N G - Machine learning videos