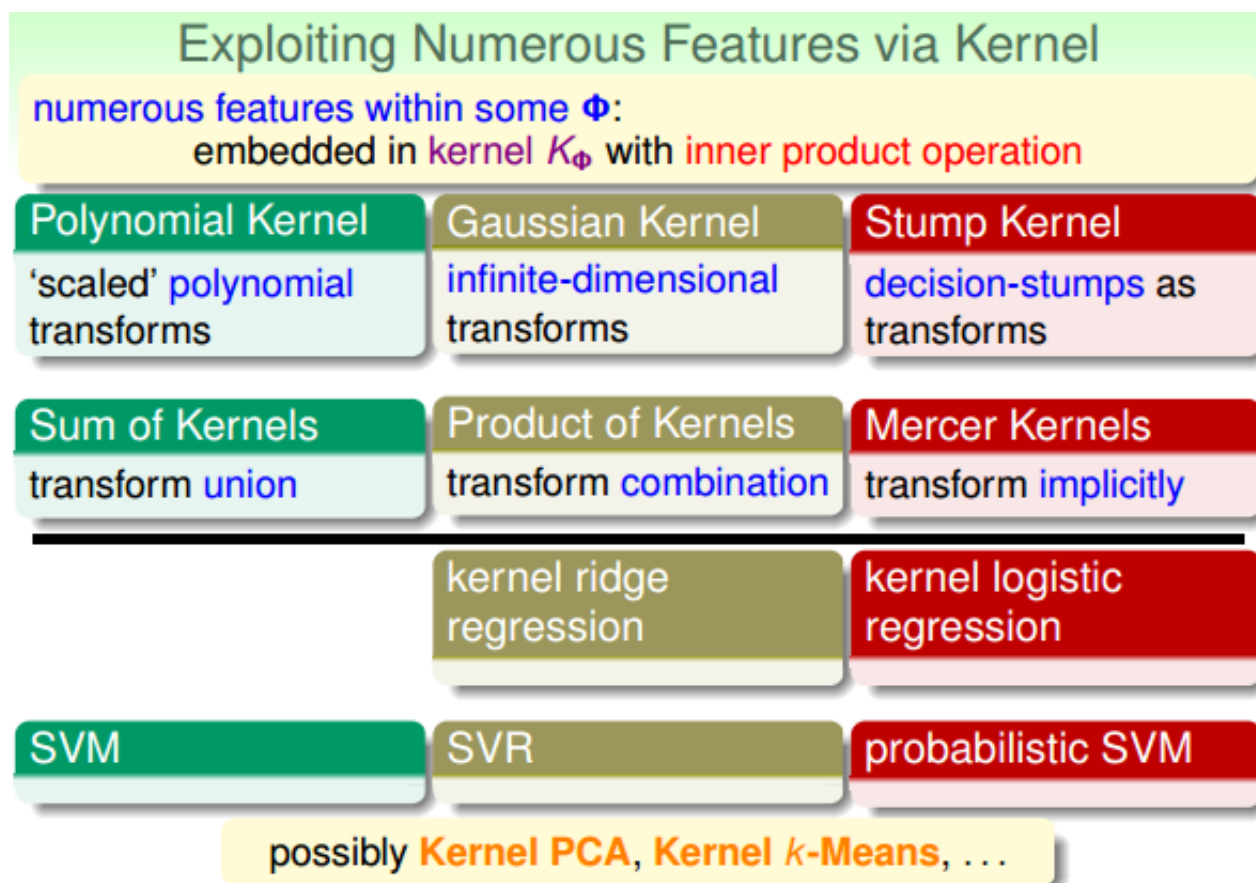# 台湾大学林轩田机器学习技法课程学习笔记16（完结） -- Finale

作者：红色石头

微信公众号：AI有道（ID：redstonewill）

上节课我们主要介绍了Matrix Factorization。通过电影推荐系统的例子，介绍Matrix Factorization其实是一个提取用户特征，关于电影的线性模型。反过来也可以看出是关于用户的线性模型。然后，我们使用SGD对模型进行最佳化。本节课我们将对机器学习技法课程介绍过的所有内容做个总结，分成三个部分：Feature Exploitation Techniques，Error Optimization Techniques和Overfitting Elimination Techniques。

Feature Exploitation Techniques

我们在本系列课程中介绍的第一个特征提取的方法就是kernel。Kernel运算将特征转换和计算内积这两个步骤合二为一，提高了计算效率。我们介绍过的kernel有：Polynormial Kernel、Gaussian Kernel、Stump Kernel等。另外，我们可以将不同的kernels相加（transform union）或者相乘（transform combination），得到不同的kernels的结合形式，让模型更加复杂。值得一提的是，要成为kernel，必须满足Mercer Condition。不同的kernel可以搭配不同的kernel模型，比如：SVM、SVR和probabilistic SVM等，还包括一些不太常用的模型：kernel ridge regression、kernel logistic regression。使用这些kernel模型就可以将线性模型扩展到非线性模型，kernel就是实现一种特征转换，从而能够处理非常复杂的非线性模型。顺便提一下，因为PCA、k-Means等算法都包含了内积运算，所以它们都对应有相应的kernel版本。

Exploiting Numerous Features via Kernel

numerous features within some Φ:
embedded in kernel K_Φ with inner product operation

Polynomial Kernel	Gaussian Kernel	Stump Kernel
'scaled' polynomial transforms	infinite-dimensional transforms	decision-stumps as transforms

Sum of Kernels	Product of Kernels	Mercer Kernels
transform union	transform combination	transform implicitly

	kernel ridge regression	kernel logistic regression
SVM	SVR	probabilistic SVM

possibly Kernel PCA, Kernel k-Means, ...

Kernel是我们利用特征转换的第一种方法，那利用特征转换的第二种方法就是Aggregation。我们之前介绍的所有的hypothesis都可以看成是一种特征转换，然后再由这些g组合成G。我们介绍过的分类模型（hypothesis）包括：Decision Stump、Decision Tree和Gaussian RBF等。如果所有的g是已知的，就可以进行blending，例如Uniform、Non-Uniform和Conditional等方式进行aggregation。如果所有的g是未知的，可以使用例如Bagging、AdaBoost和Decision Tree的方法来建立模型。除此之外，还有probabilistic SVM模型。值得一提的是，机器学习中很多模型都是类似的，我们在设计一个机器学习模型时，应该融会贯通。

Exploiting Predictive Features via Aggregation

predictive features within some Φ:
$$\phi_t(\mathbf{x}) = g_t(\mathbf{x})$$

Decision Stump	Decision Tree	(Gaussian) RBF
simplest perceptron; simplest DecTree	branching (divide) + leaves (conquer)	prototype (center) + influence

Uniform	Non-Uniform	Conditional

Bagging; Random Forest	AdaBoost; GradientBoost	Decision Tree; Nearest Neighbor

probabilistic SVM

possibly **Infinite Ensemble Learning**,
Decision Tree SVM, . . .

除此之外，我们还介绍了利用提取的方式，找出潜藏的特征（Hidden Features）。一般通过unsupervised learning的方法，从原始数据中提取出隐藏特征，使用权重表征。相应的模型包括：Neural Network、RBF Network、Matrix Factorization等。这些模型使用的unsupervised learning方法包括：AdaBoost、k-Means和Autoencoder、PCA等。

Exploiting Hidden Features via Extraction

hidden features within some **Φ**:

as hidden variables to be 'jointly' optimized with usual weights

—possibly with the help of **unsupervised learning**

Neural Network; Deep Learning	RBF Network	Matrix Factorization
neuron weights	RBF centers	user/movie factors

AdaBoost; GradientBoost	k-Means	Autoencoder; PCA
g_t parameters	cluster centers	'basis' directions

possibly **GradientBoosted Neurons**, **NNet on Factorized Features**, . . .

另外，还有一种非常有用的特征转换方法是维度压缩，即将高维度的数据降低（投影）到低维度的数据。我们介绍过的维度压缩模型包括：Decision Stump、Random Forest Tree Branching、Autoencoder、PCA和Matrix Factorization等。这些从高纬度到低纬度的特征转换在实际应用中作用很大。

Exploiting Low-Dim. Features via Compression

low-dimensional features within some **Φ**:

compressed from original features

Decision Stump; DecTree Branching	Random Forest Tree Branching	Autoencoder;PCA
'best' naïve projection to $\mathbb{R}$	'random' low-dim. projection	info.-preserving compression
		Matrix Factorization
		projection from abstract to concrete

Feature Selection

'most-helpful' low-dimensional projection

possibly **other 'dimension reduction' models**

Error Optimization Techniques

接下来我们将总结一下本系列课程中介绍过哪些优化技巧。首先，第一个数值优化技巧就是梯度下降（Gradient Descent），即让变量沿着其梯度反方向变化，不断接近最优解。例如我们介绍过的SGD、Steepest Descent和Functional GD都是利用了梯度下降的技巧。

而对于一些更复杂的最佳化问题，无法直接利用梯度下降方法来做，往往需要一些数学上的推导来得到最优解。最典型的例子是Dual SVM，还包括Kernel LogReg、Kernel RidgeReg和PCA等等。这些模型本身包含了很多数学上的一些知识，例如线性代数等等。除此之外，还有一些boosting和kernel模型，虽然本课程中没有提到，但是都会用到类似的数学推导和转换技巧。

Indirect Optimization via Equivalent Solution

when difficult to solve original problem,
seek for **equivalent solution**

Dual SVM	Kernel LogReg Kernel RidgeReg	PCA
equivalence via **convex QP**	equivalence via **representer**	equivalence to **eigenproblem**

some **other boosting models** and **modern solvers of kernel models** rely on such a technique heavily

如果原始问题比较复杂，求解比较困难，我们可以将原始问题拆分为子问题以简化计算。也就是将问题划分为多个步骤进行求解，即Multi-Stage。例如probabilistic SVM、linear blending、RBF Network等。还可以使用交叉迭代优化的方法，即Alternating Optim。例如k-Means、alternating LeastSqr等。除此之外，还可以采样分而治之的方法，即Divide & Conquer。例如decision tree。

Complicated Optimization via Multiple Steps

when difficult to solve original problem,
seek for **'easier' sub-problems**

| Multi-Stage	Alternating Optim.	Divide & Conquer
probabilistic SVM; linear blending; stacking; RBF Network; DeepNet pre-training	k-Means; alternating LeastSqr; (steepest descent)	decision tree;

useful for **complicated models**

Overfitting Elimination Techniques

Feature Exploitation Techniques和Error Optimization Techniques都是为了优化复杂

模型，减小E_{in}。但是E_{in}太小有很可能会造成过拟合overfitting。因此，机器学习中，Overfitting Elimination尤为重要。

首先，可以使用Regularization来避免过拟合现象发生。我们介绍过的方法包括：large-margin、L2、voting/averaging等等。
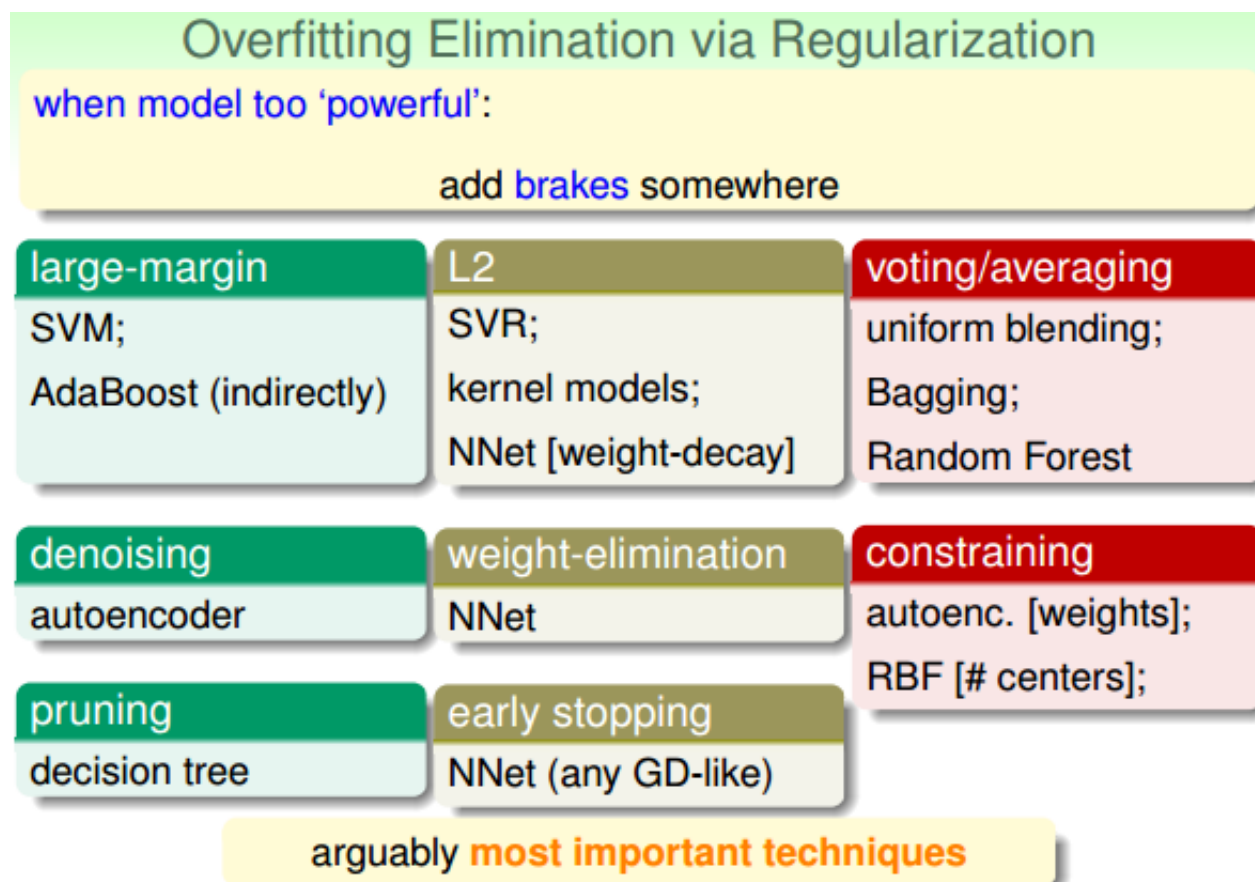
除了Regularization之外，还可以使用Validation来消除Overfitting。我们介绍过的Validation包括：SV、OOB和Internal Validation等。

Machine Learning in Action
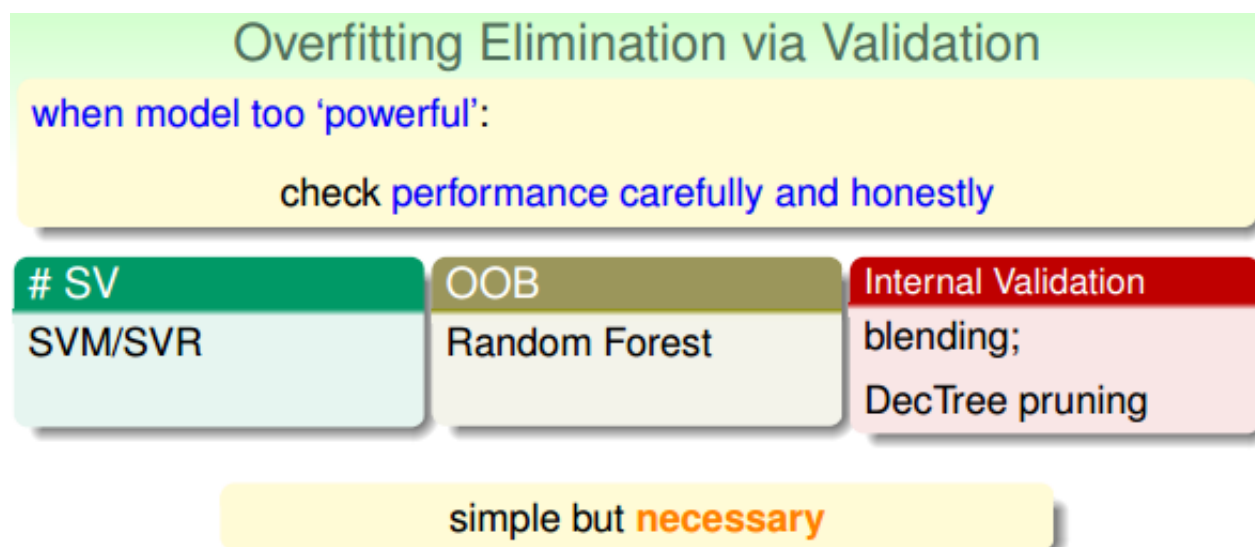
本小节介绍了林轩田老师所在的台大团队在近几年的KDDCup国际竞赛上的表现和使用的各种机器算法。融合了我们在本系列课程中所介绍的很多机器学习技法和模型。这里不再一一赘述，将相应的图片贴出来，读者自己看看吧。

NTU KDDCup 2010 World Champion Model

Feature engineering and classifier ensemble for KDD Cup 2010,
Yu et al., KDDCup 2010

linear blending of

Logistic Regression +
many rawly encoded features

Random Forest +
human-designed features

yes, you've learned everything! :-)

NTU KDDCup 2011 Track 1 World Champion Model

A linear ensemble of individual and blended models for music rating prediction,
Chen et al., KDDCup 2011

NNet, DecTree-like, and then linear blending of

- **Matrix Factorization** variants, including probabilistic **PCA**
- **Restricted Boltzmann Machines**: an 'extended' autoencoder
- *k* **Nearest Neighbors**
- **Probabilistic Latent Semantic Analysis**:
  an extraction model that has 'soft clusters' as hidden variables
- linear regression, NNet, & GBDT

**yes, you can 'easily'
understand everything! :-)**

NTU KDDCup 2012 Track 2 World Champion Model

A two-stage ensemble of diverse models for advertisement ranking in KDD Cup 2012,   Wu et al., KDDCup 2012

NNet, GBDT-like, and then linear blending of

- Linear Regression variants, including linear SVR
- Logistic Regression variants
- Matrix Factorization variants
- …

'key' is to blend properly without overfitting

NTU KDDCup 2013 Track 1 World Champion Model

Combination of feature engineering and ranking models for paper-author identification in KDD Cup 2013,   Li et al., KDDCup 2013

linear blending of

- Random Forest with many many many trees
- GBDT variants

with tons of efforts in designing features

'another key' is to construct features with domain knowledge

ICDM在2006年的时候发布了排名前十的数据挖掘算法，如下图所示。其中大部分的算法我们在本系列的课程中都有过介绍。值得一提的是Naive Bayes算法本课程中没有涉及，贝叶斯模型在实际中应用还是挺广泛的，后续可能还需要深入学习一下。

ICDM 2006 Top 10 Data Mining Algorithms

1. C4.5: another **decision tree**
2. *k*-Means
3. SVM
4. Apriori: for frequent itemset mining
5. EM: **'alternating optimization'** algorithm for some models
6. PageRank: for link-analysis, similar to **matrix factorization**
7. AdaBoost
8. *k* Nearest Neighbor
9. Naive Bayes: a simple **linear model** with 'weights' decided by data statistics
10. C&RT

personal view of five missing ML competitors:
LinReg, LogReg, Random Forest, GBDT, NNet

最后，我们将所有介绍过的机器学习算法和模型列举出来：

Machine Learning Jungle

bagging · decision tree · support vector machine · **neural network** · *kernel*

AdaBoost · aggregation · *sparsity* · autoencoder · **functional gradient**

dual · uniform blending · deep learning · nearest neighbor · decision stump

kernel LogReg · large-margin · *prototype* · quadratic programming · **SVR**

GBDT · **PCA** · random forest · *matrix factorization* · **Gaussian kernel**

soft-margin · *k-means* · OOB error · **RBF network** · probabilistic SVM

welcome to the **jungle**!

总结

本节课主要从三个方面来对机器学习技法课程做个总结：Feature Exploitation Techniques，Error Optimization Techniques和Overfitting Elimination Techniques。最后介绍了林轩田老师带领的台大团队是如何在历届KDDCup中将很多机器学习算法模型融合起来，并获得了良好的成绩。

- Feature Exploitation Techniques
  - **kernel, aggregation, extraction, low-dimensional**
- Error Optimization Techniques
  - **gradient, equivalence, stages**
- Overfitting Elimination Techniques
  - **(lots of) regularization, validation**
- Machine Learning in Practice
  - **welcome to the jungle**

注明：

文章中所有的图片均来自台湾大学林轩田《机器学习技法》课程