Some Advice on Applying Machine Learning in Practice

Yingyu Liang Computer Sciences 760 Fall 2017

http://pages.cs.wisc.edu/~yliang/cs760/

Some of the slides in these lectures have been adapted/borrowed from materials developed by Mark Craven, David Page, Jude Shavlik, Tom Mitchell, Nina Balcan, Matt Gormley, Elad Hazan, Tom Dietterich, and Pedro Domingos.

It's generalization that counts

- the fundamental goal of machine learning is generalize beyond the instances in the training set
- you should rigorously measure generalization
- use a completely held-aside test set
- or use cross validation

It's generalization that counts

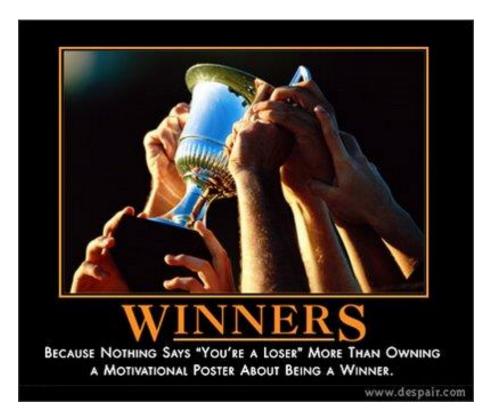
• but be careful not to let any information from test sets leak into training



• be careful about overfitting a data set, even when using cross validation

It's generalization that counts

- compare multiple learning approaches
- there is no single best approach



Data alone is not enough

- learning algorithms require inductive biases
 - smoothness
 - similar instances having similar classes
 - limited dependencies
 - limited complexity

Media Criticized For Biased Hometown Sports Reporting





CHAMPIONS! Patriots win Super Bowl, 20-17, on Vinatieri's last-second kick





By Data Shonghonney some two NEW OLLAND - This line, the hall-deletion entering associative lags, it sailed between t rights. To safet and one.

sprids to tradit and own. This time, there was its ill-disord penalty for twomany places on the look. No black cloud. No badadh. Ye Charlis From both. Adhar Visioher last owned strength and gave the Adhar Visioher last owned strength and gave the Non-Edgelan Places as 20-11 focus owner the 16.

Not a tiggined, products a port of theme call that the part here of theme at 1 hadre the 3 modern have a prober to a start of the starts of the starts of the starts backwards in a starts of theme the constraints in the problem and proves damageneously along the start backwards in a start of the start of the starts backwards in a start of the start of the starts backwards in a start of the start of the starts of the starts but to exception the maps and the of 2000 ex. The starts of the starts of the starts are starts as a start of the start of the starts and the starts of the starts of the starts of the starts and the starts of the starts of the starts of the starts and the starts of the starts of the starts. Moderney have balances in all them to starts for the starts and the starts are starts as a start of the starts. Moderney have balance as the starts of the starts of the starts and the starts and the starts are starts and the starts of the starts and starts and the starts are starts and the starts and starts are start and the starts and starts are starts and the starts are starts are starts and the starts and starts are starts and starts and starts and starts are s

weak," wall die Mr. Mawhandy terminolos, it's what happens when provide signal forms in the set of diet. And there at many provide why mit's lows." Bracht was baseling downlammed (Mill of 21, 140 panels, one transformet, we interceptions), but after 58, Looks field the panel, 17-12, with 1:56 meaning, for down for Parkins 32 panels in chipting larger (Wildow) founds.

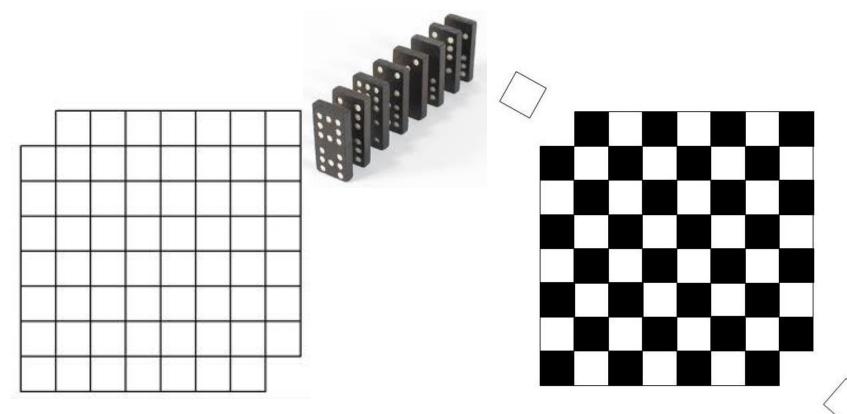
of theorem, posting Vacation in position to wise price.

Data alone is not enough

- when choosing a representation, consider what kinds of background knowledge are easily expressed in it
 - what makes instances similar \rightarrow kernels
 - dependencies \rightarrow graphical models
 - logical rules \rightarrow inductive logic programming
 - etc.

The importance of representation

- each domino covers two squares
- can you cover the board with dominoes?



• the solution is more apparent when we change the representation

Feature engineering is key

- typically the most important factor in a learning task is the feature representation
- many independent features that correlate with class \rightarrow learning is easy
- class is a complex function of features \rightarrow learning is hard
- try to craft features that make apparent what might be most important for the task

Learn many models, not just one



- winning team and runner-up were both formed by merging multiple teams
- winning systems were ensembles with > 100 models
- combination of the two winning systems was even more accurate

Learn many models, not just one

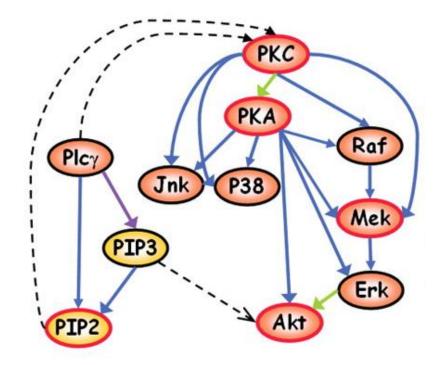
- the lesson is more general than the Netflix prize
- ensembles very often improve the accuracy of individual models

We may care more about the model than actually making predictions

- two principal reasons for using machine learning
 - 1. to make predictions about test instances
 - 2. to gain insight into the problem domain
- for the former, a complicated black box may be okay
- for the latter, we want our models to be comprehensible to some degree

We may care more about the model than actually making predictions

 example: inferring Bayesian networks to represent intracellular networks [Sachs et al., Science 2005]



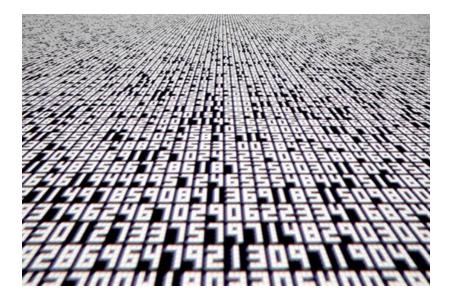
In many cases, we care about both

- example: predicting post-hospitalization VTE risk given patient histories [Kawaler et al., AMIA 2012]
 - want to identify patients at risk with high accuracy
 - want to identify previously unrecognized risk factors

Category	Risk Factor
Low Blood Volume	Furosemide
	Hypovolemia
	Hypo-osmolarity
	Posthemorrhagic Anemia
	Acute Renal Failure
Infection	E.Coli Infection
	Levofloxacin
	Cephalexin
Inflammation	High Alpha-1 Globulin Count
	Angina Pectoris
Immobilization	Pathologic Fracture of Vertebrae
Malnutrition	Protein Caloric Malnutrition

Theoretical guarantees are not what they seem

- PAC bounds are extremely loose
- asymptotic results tell us what happens when given infinite amounts of data – we don't usually have this
- learning theory results are generally
 - useful for understanding learning, driving algorithm design
 - not a criterion for practical decisions



Do assumptions of algorithm hold?

- be sure to check the assumptions made by an approach/methodology against your problem domain
 - Are the instances *i.i.d.* or should we take into account dependencies among them?
 - When we divide a data set into training/test sets, is the division representative of how the learner will be used in practice?
 - etc.
- questioning the assumptions of standard approaches sometimes results in new paradigms
 - active learning
 - multiple-instance learning
 - etc.

Compare against reasonable baselines

- Empirically determine whether fancy ML methods have value by comparing against
 - simple predictors (e.g. tomorrow's weather will be the same as today's)
 - standard predictors in use
 - individual features



