Machine Learning: Challenges and Applications

Olivier Bousquet, Pertinence
Outline

1. Motivation
2. How to do it?
3. Challenges
4. Applications
<table>
<thead>
<tr>
<th></th>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Motivation</td>
</tr>
<tr>
<td>2</td>
<td>How to do it?</td>
</tr>
<tr>
<td>3</td>
<td>Challenges</td>
</tr>
<tr>
<td>4</td>
<td>Applications</td>
</tr>
</tbody>
</table>

Olivier Bousquet, Pertinence

Machine Learning: Challenges and Applications
Spam Filtering

Motivation

How to do it?

Challenges

Applications

Olivier Bousquet, Pertinence

Machine Learning: Challenges and Applications
Re: VdAGRA

Loretta Burkitt [gaieropar@acinc.com]

Ce message a été converti en texte brut.

À: Olivier Bousquet

Hi,

VdAGRA for LESS [http://www.badesuntionkederin.com]

What is the function of this object?
Many of them.
Example 2

Re: VdIAGRA

@ Tikva Benally [kuhlmalorena@avg-joe.com]

Les sauts de ligne en surnombre de ce message ont été supprimés.
Ce message a été converti en texte brut.

À: Olivier Bousquet

Hi,

VdIAGRA for LESS http://www.kiolderunjasedfun.com

The golden ball, yes. That represents innocence, the pleasures of The woman was about my height, regal of bearing in her dark robe
Spam Filtering Algorithm

- List words that may indicate spam
- Take into account possible variations
- Count these words/variations in incoming messages
- Choose a threshold above which the message is classified as spam
- Constantly refine this algorithm as new spam gets ignored, or correct messages are rejected (update the list, change the thresholds...)

Olivier Bousquet, Pertinence

Machine Learning: Challenges and Applications
Spam Filtering Algorithm

- List words that may indicate spam
- Take into account possible variations
  - Count these words/variations in incoming messages
  - Choose a threshold above which the message is classified as spam
- Constantly refine this algorithm as new spam gets ignored, or correct messages are rejected (update the list, change the thresholds...)
Spam Filtering Algorithm

- List words that may indicate spam
- Take into account possible variations
- Count these words/variations in incoming messages
  - Choose a threshold above which the message is classified as spam
  - Constantly refine this algorithm as new spam gets ignored, or correct messages are rejected (update the list, change the thresholds...)

Olivier Bousquet, Pertinence

Machine Learning: Challenges and Applications
Spam Filtering Algorithm

- List words that may indicate spam
- Take into account possible variations
- Count these words/variations in incoming messages
- Choose a threshold above which the message is classified as spam
- Constantly refine this algorithm as new spam gets ignored, or correct messages are rejected (update the list, change the thresholds...)

Olivier Bousquet, Pertinence - Machine Learning: Challenges and Applications
Spam Filtering Algorithm

- List words that may indicate spam
- Take into account possible variations
- Count these words/variations in incoming messages
- Choose a threshold above which the message is classified as spam
- Constantly refine this algorithm as new spam gets ignored, or correct messages are rejected (update the list, change the thresholds...)

Olivier Bousquet, Pertinence
Manufacturing Process Control

- High Pressure Dye Casting

- Inject liquid aluminium in a dye, adjust temperature at various positions
- Try to avoid bubbles
Manufacturing Process Control

- High Pressure Dye Casting

- Inject liquid aluminium in a dye, adjust temperature at various positions
- Try to avoid bubbles
Manufacturing Process Control

- High Pressure Dye Casting

- Inject liquid aluminium in a dye, adjust temperature at various positions
- Try to avoid bubbles
Approach

- **Build a simulation model**
  - Use Navier-Stokes + phase transitions (solid/liquid) + air/liquid interface, complex geometry, heat transfer...
  - Require **huge** computational resources (finite elements methods)
  - **No guarantee** that the model is correct (what about impurities?)
Approach

- Build a simulation model
- Use Navier-Stokes + phase transitions (solid/liquid) + air/liquid interface, complex geometry, heat transfer...
- Require huge computational resources (finite elements methods)
- No guarantee that the model is correct (what about impurities?)
Approach

- Build a simulation model
- Use Navier-Stokes + phase transitions (solid/liquid) + air/liquid interface, complex geometry, heat transfer...
- Require huge computational resources (finite elements methods)
- No guarantee that the model is correct (what about impurities?)
Approach

- Build a simulation model
- Use Navier-Stokes + phase transitions (solid/liquid) + air/liquid interface, complex geometry, heat transfer...
- Require huge computational resources (finite elements methods)
- No guarantee that the model is correct (what about impurities?)
Is there a better way?

- Can we solve such problems in a simpler and more direct way?
- Can we build systems that can solve a wide variety of such problems (without starting from scratch for each new similar problem)?
- Idea: Use an empirical approach
Is there a better way?

- Can we solve such problems in a simpler and more direct way?
- Can we build systems that can solve a wide variety of such problems (without starting from scratch for each new similar problem)?

**Idea:** Use an empirical approach
Is there a better way?

- Can we solve such problems in a simpler and more direct way?
- Can we build systems that can solve a wide variety of such problems (without starting from scratch for each new similar problem)?
- Idea: Use an empirical approach
Outline

1. Motivation
2. How to do it?
3. Challenges
4. Applications
Induction

- **Start from Data or Observations**
- Build a model which
  - Agrees with the data
  - Predicts unobserved data
- **Scientific Method**: laws are induced from observations. A law is correct as long as no experiment contradicts it.
Induction

- Start from Data or Observations
- Build a model which
  - Agrees with the data
  - Predicts unobserved data
- Scientific Method: laws are induced from observations. A law is correct as long as no experiment contradicts it.
Induction

- Start from Data or Observations
- Build a model which
  - Agrees with the data
  - Predicts unobserved data
- Scientific Method: laws are induced from observations. A law is correct as long as no experiment contradicts it.
Induction

- Start from Data or Observations
- Build a model which
  - Agrees with the data
  - Predicts unobserved data
- Scientific Method: laws are induced from observations. A law is correct as long as no experiment contradicts it.
Induction

- Start from Data or Observations
- Build a model which
  - Agrees with the data
  - Predicts unobserved data
- Scientific Method: laws are induced from observations. A law is correct as long as no experiment contradicts it.
Ingredients 1

- Data representation: choose a way to represent the objects to be classified
Example

Olivier Bousquet, Pertinence

Machine Learning: Challenges and Applications
Example

1 0 1 0 0
Ingredients 2

- Data representation: choose a way to represent the objects to be classified
- Class of functions: choose a way to express the classification function
Example
Data is not always nicely separated
Overfitting

Underfitting.
Overfitting
Overfitting

Reasonable model.
Ingredients

- Data representation: choose a way to represent the objects to be classified
- Class of functions: choose a way to express the classification function
- Preference: incorporate assumptions about the model regularity
- Algorithm: set up the problem as an optimization problem
Ingredients

- Data representation: choose a way to represent the objects to be classified
- Class of functions: choose a way to express the classification function
- Preference: incorporate assumptions about the model regularity
- Algorithm: set up the problem as an optimization problem
Ingredients

- Data representation: choose a way to represent the objects to be classified
- Class of functions: choose a way to express the classification function
- Preference: incorporate assumptions about the model regularity
- Algorithm: set up the problem as an optimization problem
Formalization

Example: statistical formalization (not the only one)

- Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)
- Function class: \(\mathcal{F}, f : \mathcal{X} \rightarrow \mathcal{Y}\)
- Loss function: \(\ell(f(x), y) = 1_{[f(x) \neq y]}\)
- Goal: find a function \(f \in \mathcal{F}\) such that \(\mathbb{E}\ell(f(X), Y)\) is as small as possible, and \(f\) is regular enough
- Optimization problem

\[
\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(X_i), Y_i) + \lambda \|f\| \]

- Key question: convergence of \(\mathbb{E}\ell(\hat{f}(X), Y)\) to \(\min_{f} \mathbb{E}\ell(f(X), Y)\)
Formalization

Example: statistical formalization (not the only one)

- Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)
- Function class: \(\mathcal{F}, f : \mathcal{X} \rightarrow \mathcal{Y}\)
- Loss function: \(\ell(f(x), y) = 1_{[f(x) \neq y]}\)
- Goal: find a function \(f \in \mathcal{F}\) such that \(\mathbb{E}\ell(f(X), Y)\) is as small as possible, and \(f\) is regular enough
- Optimization problem

\[
\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(X_i), Y_i) + \lambda \|f\|
\]

- Key question: convergence of \(\mathbb{E}\ell(\hat{f}(X), Y)\) to \(\min_{f} \mathbb{E}\ell(f(X), Y)\)
Formalization

Example: statistical formalization (not the only one)

- Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)
- Function class: \(\mathcal{F}, f : \mathcal{X} \rightarrow \mathcal{Y}\)
- Loss function: \(\ell(f(x), y) = 1_{f(x) \neq y}\)
- Goal: find a function \(f \in \mathcal{F}\) such that \(\mathbb{E}\ell(f(X), Y)\) is as small as possible, and \(f\) is regular enough
- Optimization problem

\[
\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(X_i), Y_i) + \lambda \|f\|
\]

- Key question: convergence of \(\mathbb{E}\ell(\hat{f}(X), Y)\) to \(\min_{f} \mathbb{E}\ell(f(X), Y)\)
Formalization

Example: statistical formalization (not the only one)

- Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)
- Function class: \(\mathcal{F}, f : \mathcal{X} \to \mathcal{Y}\)
- Loss function: \(\ell(f(x), y) = 1[f(x) \neq y]\)
- Goal: find a function \(f \in \mathcal{F}\) such that \(\mathbb{E}\ell(f(X), Y)\) is as small as possible, and \(f\) is regular enough

Optimization problem

\[
\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(X_i), Y_i) + \lambda \|f\|
\]

- Key question: convergence of \(\mathbb{E}\ell(\hat{f}(X), Y)\) to \(\min_{f} \mathbb{E}\ell(f(X), Y)\)
Formalization

Example: statistical formalization (not the only one)

- Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)
- Function class: \(\mathcal{F}, f : \mathcal{X} \to \mathcal{Y}\)
- Loss function: \(\ell(f(x), y) = 1_{f(x) \neq y}\)
- Goal: find a function \(f \in \mathcal{F}\) such that \(\mathbb{E}\ell(f(X), Y)\) is as small as possible, and \(f\) is regular enough

Optimization problem

\[
\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(X_i), Y_i) + \lambda \|f\|
\]

- Key question: convergence of \(\mathbb{E}\ell(\hat{f}(X), Y)\) to \(\min_{f} \mathbb{E}\ell(f(X), Y)\)
Formalization

Example: statistical formalization (not the only one)

- Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)
- Function class: \(\mathcal{F}, f : \mathcal{X} \rightarrow \mathcal{Y}\)
- Loss function: \(\ell(f(x), y) = 1[f(x) \neq y]\)
- Goal: find a function \(f \in \mathcal{F}\) such that \(\mathbb{E}\ell(f(X), Y)\) is as small as possible, and \(f\) is regular enough
- Optimization problem

\[
\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(X_i), Y_i) + \lambda \|f\|
\]

- Key question: convergence of \(\mathbb{E}\ell(\hat{f}(X), Y)\) to \(\min_{f} \mathbb{E}\ell(f(X), Y)\)
Outline

1 Motivation

2 How to do it?

3 Challenges

4 Applications
High Dimensionality

- Complexity of the data increases: requires bigger description vectors
- Gathering data becomes easier: more and more descriptors available
- This increase in the dimensionality comes at a price
  - Overfitting becomes the main issue (curse of dimensionality)
  - But at the same time, interesting phenomena occur (blessing)
High Dimensionality

- Complexity of the data increases: requires bigger description vectors
- Gathering data becomes easier: more and more descriptors available
- This increase in the dimensionality comes at a price
  - Overfitting becomes the main issue (curse of dimensionality)
  - But at the same time, interesting phenomena occur (blessing)
High Dimensionality

- Complexity of the data increases: requires bigger description vectors
- Gathering data becomes easier: more and more descriptors available
- This increase in the dimensionality comes at a price
  - Overfitting becomes the main issue (curse of dimensionality)
  - But at the same time, interesting phenomena occur (blessing)
Curses

- NP-completeness: optimization problems can be exponential in the dimension (e.g. search for the separating hyperplane with minimum number of mistakes)
- Statistical issue: exponentially slow convergence

Theorem

Let $\mathcal{F}$ be the set of all Lipschitz functions on $[0, 1]^d$. For any estimator,

$$\sup_{f \in \mathcal{F}} \mathbb{E}(\hat{f}(x) - f(x))^2 \geq Cn^{-\frac{2}{2+d}} \text{ as } n \to \infty$$
NP-completeness: optimization problems can be exponential in the dimension (e.g. search for the separating hyperplane with minimum number of mistakes)

Statistical issue: exponentially slow convergence

Theorem

Let $\mathcal{F}$ be the set of all Lipschitz functions on $[0, 1]^d$. For any estimator,

$$
\sup_{f \in \mathcal{F}} \mathbb{E}((\hat{f}(x) - f(x))^2) \geq Cn^{-\frac{2}{2+d}} \text{ as } n \to \infty
$$
Concentration-of-measure phenomenon

- On the $n$-sphere, for the usual rotationally-invariant measure, most of the mass is concentrated near any $(n-1)$-sphere that is equatorial in it (analogous to a great circle on the Earth’s surface, when $n = 2$). In other words, the ’poles’ and their neighborhoods down to very small latitudes account for a tiny proportion of the ’area’.

- Chernoff bounds: for i.i.d. bounded random variables

\[
P\left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}X \geq \epsilon\right) \leq e^{-nc\epsilon^2}
\]

- General concentration inequality: under appropriate conditions on $f$ (e.g. Lipschitz)

\[
P[f(X_1, \ldots, X_n) - \mathbb{E}f \geq \epsilon] \leq e^{-nc\epsilon^2}
\]
Blessings

Concentration-of-measure phenomenon

- On the $n$-sphere, for the usual rotationally-invariant measure, most of the mass is concentrated near any $(n-1)$-sphere that is equatorial in it (analogous to a great circle on the Earth’s surface, when $n = 2$). In other words, the 'poles' and their neighborhoods down to very small latitudes account for a tiny proportion of the 'area'.

- Chernoff bounds: for i.i.d. bounded random variables

\[
P\left[ \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}X \geq \epsilon \right] \leq e^{-n\epsilon^2}
\]

- General concentration inequality: under appropriate conditions on $f$ (e.g. Lipschitz)

\[
P\left[ f(X_1, \ldots, X_n) - \mathbb{E}f \geq \epsilon \right] \leq e^{-n\epsilon^2}
\]
Motivation

How to do it?

Challenges

Applications

Blessings

Concentration-of-measure phenomenon

- On the \( n \)-sphere, for the usual rotationally-invariant measure, most of the mass is **concentrated** near any \((n - 1)\)-sphere that is equatorial in it (analogous to a great circle on the Earth’s surface, when \( n = 2 \)). In other words, the 'poles' and their neighborhoods down to very small latitudes account for a tiny proportion of the 'area'.

- Chernoff bounds: for i.i.d. bounded random variables

\[
\mathbb{P} \left[ \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}X \geq \epsilon \right] \leq e^{-nc\epsilon^2}
\]

- General concentration inequality: under appropriate conditions on \( f \) (e.g. Lipschitz)

\[
\mathbb{P} [f(X_1, \ldots, X_n) - \mathbb{E}f \geq \epsilon] \leq e^{-nc\epsilon^2}
\]
Random Projections

**Lemma (Johnson & Lindenstrauss)**

*Given a set $S$ of points in $\mathbb{R}^d$, if we perform a random orthogonal projection of those points on a subspace of dimension $m$, then $m = O(\gamma^{-2} \log |S|)$ is sufficient so that with high probability all pairwise distances are preserved up to a factor $1 \pm \gamma$*

→ cheap and easy way to reduce the dimension, while preserving the geometry
Connections

- **Theoretical**
  - Logic
  - Information Theory / Compression
  - Algorithmic Information Theory
  - Mathematical Statistics

- **High dimensional phenomena**
  - Probability in Banach spaces
  - Random graphs, graph theory
  - Concentration

- **Algorithmics**
  - Optimization
  - Approximate algorithms
Connections

- Theoretical
  - Logic
  - Information Theory / Compression
  - Algorithmic Information Theory
  - Mathematical Statistics
- High dimensional phenomena
  - Probability in Banach spaces
  - Random graphs, graph theory
  - Concentration
- Algorithmics
  - Optimization
  - Approximate algorithms
Connections

- **Theoretical**
  - Logic
  - Information Theory / Compression
  - Algorithmic Information Theory
  - Mathematical Statistics

- **High dimensional phenomena**
  - Probability in Banach spaces
  - Random graphs, graph theory
  - Concentration

- **Algorithmics**
  - Optimization
  - Approximate algorithms
Dealing with Dimensionality

\[
\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(X_i), Y_i) + \lambda \|f\| 
\]

Important questions

- How to choose the norm?
- What can be said about convergence?
- How can this be implemented?
Dealing with Dimensionality

\[
\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(X_i), Y_i) + \lambda \|f\|
\]

Important questions

- How to choose the norm?
- What can be said about convergence?
- How can this be implemented?
Dealing with Dimensionality

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(X_i), Y_i) + \lambda \|f\|$$

Important questions

- How to choose the norm?
- What can be said about convergence?
- How can this be implemented?
Dealing with Dimensionality

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(X_i), Y_i) + \lambda \|f\|$$

Important questions

- How to choose the norm?
- What can be said about convergence?
- How can this be implemented?
Choosing the norm

- Typical approach: linear methods
  \[ f(x) = \sum_{i=1}^{d} \alpha_i x^i \]

- \( L_2 \) norm:
  - by duality everything can be expressed in terms of inner products
  \[ \sum_{i=1}^{d} x^i z^i \]
  - this allows to generalize to reproducing kernel Hilbert spaces

- \( L_1 \): ensures sparsity
  - only a few \( \alpha_i \) will be non-zero
  - this allows to deal with very high-dimensional representations
Choosing the norm

- Typical approach: linear methods

\[ f(x) = \sum_{i=1}^{d} \alpha_i x_i \]

- \( L_2 \) norm:
  - by duality everything can be expressed in terms of inner products
  \[ \sum_{i=1}^{d} x_i z_i \]
  - this allows to generalize to reproducing kernel Hilbert spaces
- \( L_1 \): ensures sparsity
  - only a few \( \alpha_i \) will be non-zero
  - this allows to deal with very high-dimensional representations
Choosing the norm

- Typical approach: linear methods

\[ f(x) = \sum_{i=1}^{d} \alpha_i x^i \]

- \( L_2 \) norm:
  - by duality everything can be expressed in terms of inner products

\[ \sum_{i=1}^{d} x^i z^i \]

- this allows to generalize to reproducing kernel Hilbert spaces

- \( L_1 \): ensures sparsity
  - only a few \( \alpha_i \) will be non-zero
  - this allows to deal with very high-dimensional representations
Choosing the norm

- Typical approach: linear methods

\[ f(x) = \sum_{i=1}^{d} \alpha_i x^i \]

- \( L_2 \) norm:
  - by duality everything can be expressed in terms of inner products
  \[ \sum_{i=1}^{d} x^i z^i \]
  - this allows to generalize to reproducing kernel Hilbert spaces

- \( L_1 \): ensures sparsity
  - only a few \( \alpha_i \) will be non-zero
  - this allows to deal with very high-dimensional representations
Choosing the norm

- Typical approach: linear methods

\[ f(x) = \sum_{i=1}^{d} \alpha_i x^i \]

- \(L_2\) norm:
  - by duality everything can be expressed in terms of inner products

\[ \sum_{i=1}^{d} x^i z^i \]

  - this allows to generalize to reproducing kernel Hilbert spaces

- \(L_1\): ensures sparsity
  - only a few \(\alpha_i\) will be non-zero
  - this allows to deal with very high-dimensional representations
Choosing the norm

- Typical approach: linear methods

\[ f(x) = \sum_{i=1}^{d} \alpha_i x^i \]

- \( L_2 \) norm:
  - by duality everything can be expressed in terms of inner products

\[ \sum_{i=1}^{d} x^i z^i \]

  - this allows to generalize to reproducing kernel Hilbert spaces

- \( L_1 \): ensures sparsity
  - only a few \( \alpha_i \) will be non-zero
  - this allows to deal with very high-dimensional representations
Choosing the norm

- Typical approach: linear methods
  \[ f(x) = \sum_{i=1}^{d} \alpha_i x^i \]

- \( L_2 \) norm:
  - by duality everything can be expressed in terms of inner products
  \[ \sum_{i=1}^{d} x^i z^i \]
  - this allows to generalize to reproducing kernel Hilbert spaces

- \( L_1 \): ensures sparsity
  - only a few \( \alpha_i \) will be non-zero
  - this allows to deal with very high-dimensional representations
**L₂: Kernels**

- **Idea:** replace inner products by positive definite kernels

- Three reasons why this is interesting:
  - a good way to measure smoothness in high dimensions
  - allows to deal with complex objects
  - easy way to non-linearize

- Also, interesting geometric properties (balls are effectively low dimensional in $L₂(P)$)
Motivation
How to do it?
Challenges
Applications

$L_2$: Kernels

- Idea: replace inner products by positive definite kernels
- Three reasons why this is interesting:
  - a good way to measure smoothness in high dimensions
  - allows to deal with complex objects
  - easy way to non-linearize
- Also, interesting geometric properties (balls are effectively low dimensional in $L_2(P)$)
L₂: Kernels

- Idea: replace inner products by positive definite kernels
- Three reasons why this is interesting:
  - a good way to measure smoothness in high dimensions
  - allows to deal with complex objects
  - easy way to non-linearize
- Also, interesting geometric properties (balls are effectively low dimensional in L₂(ℕ))
### L₁: Boosting

- Use sparsity to introduce many dimensions
- Given a set $\mathcal{H}$ of *interesting* features, replace $x$ by $(h(x))_{h \in \mathcal{H}}$
- Build a linear combination with small $L_1$ norm
- Effective algorithms (e.g. Adaboost), geometry not significantly altered by replacing $\mathcal{H}$ by its convex hull
$L_1$: Boosting

- Use sparsity to introduce many dimensions
- Given a set $\mathcal{H}$ of *interesting* features, replace $x$ by $(h(x))_{h \in \mathcal{H}}$
- Build a linear combination with small $L_1$ norm
- Effective algorithms (e.g., Adaboost), geometry not significantly altered by replacing $\mathcal{H}$ by its convex hull
### $L_1$: Boosting

- Use sparsity to introduce many dimensions
- Given a set $\mathcal{H}$ of *interesting* features, replace $x$ by $(h(x))_{h \in \mathcal{H}}$
- Build a linear combination with small $L_1$ norm
- Effective algorithms (e.g. Adaboost), geometry not significantly altered by replacing $\mathcal{H}$ by its convex hull
Motivation

How to do it?

Challenges

Applications

L₁: Boosting

- Use sparsity to introduce many dimensions
- Given a set \( \mathcal{H} \) of *interesting* features, replace \( x \) by \((h(x))_{h \in \mathcal{H}}\)
- Build a linear combination with small \( L_1 \) norm
- Effective algorithms (e.g. Adaboost), geometry not significantly altered by replacing \( \mathcal{H} \) by its convex hull
Convergence Properties

- Concentration ensures

\[
\sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) - \mathbb{E}\ell(f(X), Y) \approx \mathbb{E} \left[ \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) - \mathbb{E}\ell(f(X), Y) \right]
\]

- Expectation term can be written as \( \mathbb{E}\|\sum X_i\| \)

- Key concept: geometry induced by \( F \) and the distribution, "size" of \( F \) can be measured by metric entropy, Rademacher averages, majorizing measures,...
Convergence Properties

- Concentration ensures:
  \[
  \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) - \mathbb{E}\ell(f(X), Y) \approx \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) - \mathbb{E}\ell(f(X), Y) \right]
  \]

- Expectation term can be written as \( \mathbb{E}\| \sum X_i \| \)

- Key concept: geometry induced by \( \mathcal{F} \) and the distribution, ”size” of \( \mathcal{F} \) can be measured by metric entropy, Rademacher averages, majorizing measures, ...
Convergence Properties

- Concentration ensures

\[
\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) - \mathbb{E} \ell(f(X), Y) \approx \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) - \mathbb{E} \right]
\]

- Expectation term can be written as \( \mathbb{E} \| \sum X_i \| \)

- Key concept: geometry induced by \( \mathcal{F} \) and the distribution, "size" of \( \mathcal{F} \) can be measured by metric entropy, Rademacher averages, majorizing measures, ...
Typically use convex optimization (but number of variables can be huge)

- Kernels: fast way of computing inner products
- Boosting: few relevant features, fast selection of most important feature
- Random Projections: cheap and easy way to reduce dimensionality
Computational Tricks

Typically use convex optimization (but number of variables can be huge)

- Kernels: fast way of computing inner products
- Boosting: few relevant features, fast selection of most important feature
- Random Projections: cheap and easy way to reduce dimensionality
Computational Tricks

Typically use convex optimization (but number of variables can be huge)

- Kernels: fast way of computing inner products
- Boosting: few relevant features, fast selection of most important feature
- Random Projections: cheap and easy way to reduce dimensionality
Typically use convex optimization (but number of variables can be huge)

- Kernels: fast way of computing inner products
- Boosting: few relevant features, fast selection of most important feature
- Random Projections: cheap and easy way to reduce dimensionality
Outline

1. Motivation
2. How to do it?
3. Challenges
4. Applications
To be discussed next

- A wide variety of applications
- Future directions
To be discussed next

- A wide variety of applications
- Future directions
**Recommendation Engines**

**Motivation**

- How to do it?
- Challenges
- Applications

**Olivier Bousquet, Pertinence**

**Word of Mouth Marketing: How Smart Companies Get People Talking (Hardcover)**

by Andy Sernotiz, Seth Godin (Foreword), Guy Kawasaki (Afterword)

- List Price: $24.96
- Prices: $16.47 & eligible for FREE Super Saver Shipping on orders over $25. Details
- You Save: $8.49 (34%)

Pre-Order Price Guaranteed! Order now and if the Amazon.com price decreases between your order time and release date, you'll receive the lowest price. See Details

Availability: This title will be released on November 1, 2006. Pre-order now. Ships from and sold by Amazon.com.

**Better Together**

Buy this book with The Long Tail: Why the Future of Business is Selling Less of More by Chris Anderson to Buy Together Today: $32.94

Customers who bought this item also bought
- The Long Tail: Why the Future of Business is Selling Less of More by Chris Anderson
- Small Is the New Big: and 102 Other Riffs, Rants, and Remarkable Business Ideas by Seth Godin
News Categorization

Des actionnaires veulent réviser la fusion
Nouvel Observateur - Oct 20, 2006
Des actionnaires de Suez veulent réviser la parité de fusion, en demandant à GDF le versement d'un dividende majoré. La direction de Suez a reçu une lettre de plusieurs actionnaires étrangers demandant ...

Deux semaines cruciales pour la fusion Suez-Gaz de France
Fusion : des actionnaires de Suez réclament une revalorisation ...

Le Figaro - Investir.fr
all 23 news articles »
Des actionnaires veulent réviser la fusion

NOUVELOBS.COM | 20.10.06 | 16:44

Des actionnaires de Suez veulent réviser la parité de fusion, en demandant à GDF le versement d'un dividende majoré.

La direction de Suez a reçu une lettre de plusieurs actionnaires étrangers demandant une revalorisation des termes de sa fusion avec Gaz de France. L'information, publiée sur le site du Wall Street Journal, a été confirmée, vendredi 20 octobre, par une porte-parole du groupe. "Je confirme que nous avons reçu cette lettre..."
Deux semaines cruciales pour la fusion Suez-Gaz de France

Par William Emmanuel

PARIS (Reuters) - Suez et Gaz de France n’ont plus que deux semaines pour régler leurs différends s’ils veulent boucler leur projet de fusion avant la fin de l’année mais un échec n’est plus exclu. /Photo d’archives/REUTERS/Charles Platiau (cliquez pour agrandir)

Les tensions apparues entre les deux groupes sur la répartition des postes de direction font en effet craindre un report qui pourrait être fatal, l’approche de l’élection présidentielle d’avril constituant un obstacle de taille.

Si l’opération ne se faisait pas, les analystes financiers et des responsables industriels pensent que le groupe d’énergie et de services à l’environnement serait rapidement la cible d’une OPA d’un grand concurrent européen et serait démantelé.
**Fusion : des actionnaires de Suez réclament une revalorisation de l'échange**

LE MONDE | 21.10.06 | 15h16 • Mis à jour le 21.10.06 | 15h17

Consultez les dossiers, l'analyse approfondie des grands sujets d'actualité.
Abonnez-vous au Monde.fr : 6€ par mois + 30 jours offerts

Juste au bout, le projet de mariage entre Suez et Gaz de France (GDF) aura été contrarié. Par les parlementaires de gauche, qui exigent plus d'Etat. Par la Commission européenne et le gouvernement belge, qui réclament plus de concurrence.

Et aujourd'hui par les actionnaires de Suez, qui veulent plus d'argent. A un peu plus de deux mois de leur assemblée générale, décisive pour la fusion, une partie d'entre eux réaffirme son hostilité à l'opération et cherche à fédérer les mécontents.
Autonomous Vehicles

Motivation

How to do it

Challenges

Applications

Olivier Bousquet, Pertinence

Machine Learning: Challenges and Applications
Other Applications

- Bioinformatics
- Recognition problems (speech, handwritten text, images...)
- Data Mining
- Search Engines
- ...

Olivier Bousquet, Pertinence
Other Applications

- Bioinformatics
- Recognition problems (speech, handwritten text, images...)
- Data Mining
- Search Engines
- ...

Olivier Bousquet, Pertinence
Other Applications

- Bioinformatics
- Recognition problems (speech, handwritten text, images...)
- Data Mining
  - Search Engines
- ...
Other Applications

- Bioinformatics
- Recognition problems (speech, handwritten text, images...)
- Data Mining
- Search Engines
- ...

Olivier Bousquet, Pertinence

Machine Learning: Challenges and Applications
Other Applications

- Bioinformatics
- Recognition problems (speech, handwritten text, images...)
- Data Mining
- Search Engines
- ...

Olivier Bousquet, Pertinence
Machine Learning Can Make You Rich

- Mathematical Finance is the scientific topic of choice for getting a good job
- Machine Learning is not too bad either these days!
- Recruiting companies: big shots of the internet betting heavily on this technology
  - Google / Yahoo! / Microsoft
  - Also some start-ups: Pertinence
- Money Prizes
  - DARPA Grand Challenge 2005: autonomous vehicle $2M (http://www.grandchallenge.org)
  - Netflix Prize 2006: movie ratings $1M (http://www.netflixprize.com)
  - Hutter Prize: compression of web documents $50K (http://prize.hutter1.net/)
Machine Learning Can Make You Rich

- **Motivation**
  - Mathematical Finance is the scientific topic of choice for getting a good job
  - Machine Learning is not too bad either these days!
  - Recruiting companies: big shots of the internet betting heavily on this technology
    - Google / Yahoo! / Microsoft
    - Also some start-ups: Pertinence
  - Money Prizes
    - DARPA Grand Challenge 2005: autonomous vehicle $2M
      (http://www.grandchallenge.org)
    - Netflix Prize 2006: movie ratings $1M
      (http://www.netflixprize.com)
    - Hutter Prize: compression of web documents $50K
      (http://prize.hutter1.net/)
Machine Learning Can Make You Rich

- Mathematical Finance is the scientific topic of choice for getting a good job
- Machine Learning is not too bad either these days!
- Recruiting companies: big shots of the internet betting heavily on this technology
  - Google / Yahoo! / Microsoft
  - Also some start-ups: Pertinence
- Money Prizes
  - DARPA Grand Challenge 2005: autonomous vehicle $ 2M
    ([http://www.grandchallenge.org](http://www.grandchallenge.org))
  - Netflix Prize 2006: movie ratings $ 1M
    ([http://www.netflixprize.com](http://www.netflixprize.com))
  - Hutter Prize: compression of web documents $ 50K
    ([http://prize.hutter1.net/](http://prize.hutter1.net/))
Machine Learning Can Make You Rich

- Mathematical Finance is the scientific topic of choice for getting a good job
- Machine Learning is not too bad either these days!
- Recruiting companies: big shots of the internet betting heavily on this technology
  - Google / Yahoo! / Microsoft
  - Also some start-ups: Pertinence

- Money Prizes
  - DARPA Grand Challenge 2005: autonomous vehicle $ 2M
    (http://www.grandchallenge.org)
  - Netflix Prize 2006: movie ratings $ 1M
    (http://www.netflixprize.com)
  - Hutter Prize: compression of web documents $ 50K
    (http://prize.hutter1.net/)
Machine Learning Can Make You Rich

- Mathematical Finance is the scientific topic of choice for getting a good job
- Machine Learning is not too bad either these days!
- Recruiting companies: big shots of the internet betting heavily on this technology
  - Google / Yahoo! / Microsoft
  - Also some start-ups: Pertinence

- Money Prizes
  - DARPA Grand Challenge 2005: autonomous vehicle $ 2M
    (http://www.grandchallenge.org)
  - Netflix Prize 2006: movie ratings $ 1M
    (http://www.netflixprize.com)
  - Hutter Prize: compression of web documents $ 50K
    (http://prize.hutter1.net/)
Machine Learning Can Make You Rich

- Mathematical Finance is the scientific topic of choice for getting a good job
- Machine Learning is not too bad either these days!
- Recruiting companies: big shots of the internet betting heavily on this technology
  - Google / Yahoo! / Microsoft
  - Also some start-ups: Pertinence
- Money Prizes
  - DARPA Grand Challenge 2005: autonomous vehicle $2M (http://www.grandchallenge.org)
  - Netflix Prize 2006: movie ratings $1M (http://www.netflixprize.com)
  - Hutter Prize: compression of web documents $50K (http://prize.hutter1.net/)
Future Directions

- Practically
  - Massive databases: require specific algorithms
  - More and more structure: e.g. machine translation
  - Multimedia: e.g. video streams

- Theoretically
  - Better connection between approximation/estimation/computation
  - Better understanding of model selection (e.g. cross-validation)
  - Formalization and understanding of feature selection