Learning a set of directions

Wouter M. Koolen Jiazhong Nie Manfred K. Warmuth

Outline

Motivation

Measuring gain

Algorithm

Conclusion

Problem

Parts of my home town Amsterdam lie 5 metres below sea level

☹

Solution

Pump out water

Leeghwater (1607)

This is how we do it

And then global warming sets in . . .

Online learning to the rescue

For $t = 1, 2, \ldots$

- Mill chooses a direction $\boldsymbol{u}_t$
- Wind reveals direction $\boldsymbol{x}_t$
- Gain based on match

What is a reasonable gain?

Measuring gain

Gain quantifies quality of prediction u upon outcome x

Measuring gain

Gain quantifies quality of prediction $\boldsymbol{u}$ upon outcome $\boldsymbol{x}$

Perhaps the simplest gain is the

$$\textbf{angle cosine} \;:=\; \boldsymbol{u}^\intercal \boldsymbol{x}$$

Measuring gain

Gain quantifies quality of prediction $\boldsymbol{u}$ upon outcome $\boldsymbol{x}$

Perhaps the simplest gain is the

$$\textbf{angle cosine} := \boldsymbol{u}^\intercal \boldsymbol{x}$$

best when $\boldsymbol{u}$,$\boldsymbol{x}$ parallel, worst when $\boldsymbol{u}$,$\boldsymbol{x}$ opposite

Measuring gain

Gain quantifies quality of prediction $\boldsymbol{u}$ upon outcome $\boldsymbol{x}$

Perhaps the simplest gain is the

$$\textbf{angle cosine} \;:=\; \boldsymbol{u}^\intercal \boldsymbol{x}$$

best when $\boldsymbol{u}$,$\boldsymbol{x}$ parallel, worst when $\boldsymbol{u}$,$\boldsymbol{x}$ opposite

Another gain is used in Principal Component Analysis

$$\textbf{subspace similarity} \;:=\; \left(\boldsymbol{u}^\intercal \boldsymbol{x}\right)^2$$

Measuring gain

Gain quantifies quality of prediction $\boldsymbol{u}$ upon outcome $\boldsymbol{x}$

Perhaps the simplest gain is the

$$\textbf{angle cosine} \;:=\; \boldsymbol{u}^\intercal \boldsymbol{x}$$

best when $\boldsymbol{u}$,$\boldsymbol{x}$ parallel, worst when $\boldsymbol{u}$,$\boldsymbol{x}$ opposite

Another gain is used in Principal Component Analysis

$$\textbf{subspace similarity} \;:=\; \left(\boldsymbol{u}^\intercal \boldsymbol{x}\right)^2$$

best when $\boldsymbol{u}$,$\boldsymbol{x}$ parallel or opposite

Measuring gain

Gain quantifies quality of prediction $\boldsymbol{u}$ upon outcome $\boldsymbol{x}$

Perhaps the simplest gain is the

$$\textbf{angle cosine} \ := \ \boldsymbol{u}^\intercal \boldsymbol{x}$$

best when $\boldsymbol{u}$,$\boldsymbol{x}$ parallel, worst when $\boldsymbol{u}$,$\boldsymbol{x}$ opposite

Another gain is used in Principal Component Analysis

$$\textbf{subspace similarity} \ := \ \left(\boldsymbol{u}^\intercal \boldsymbol{x}\right)^2$$

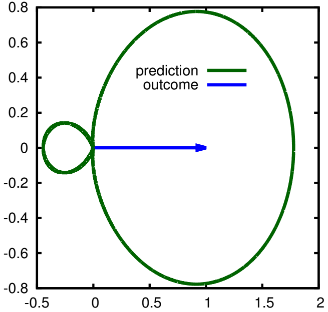
best when $\boldsymbol{u}$,$\boldsymbol{x}$ parallel or opposite

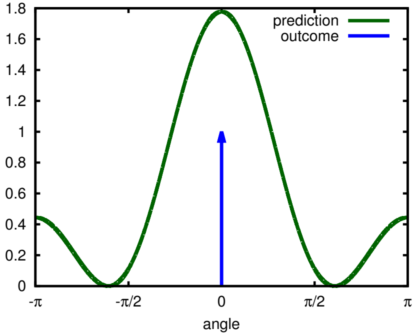
Our solution: controlled trade-off (windmill-dependent constant c)

$$\textbf{directional gain} \ := \ \left(\boldsymbol{u}^\intercal \boldsymbol{x} + c\right)^2$$

Visualisation of directional gain

$$\left(\boldsymbol{u}^\intercal \boldsymbol{x} + c\right)^2 \qquad \text{with } c = 1/3$$





Gain expansion

For randomised prediction $\boldsymbol{u} \sim \mathbb{P}$:

$$\begin{aligned}\mathbb{E}\left[\left(\boldsymbol{u}^\intercal \boldsymbol{x} + c\right)^2\right] &= \mathbb{E}\left[\boldsymbol{x}^\intercal \boldsymbol{u}\boldsymbol{u}^\intercal \boldsymbol{x} + 2c\boldsymbol{x}^\intercal \boldsymbol{u} + c^2\right] \\ &= \boldsymbol{x}^\intercal \mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}^\intercal\right]\boldsymbol{x} + 2c\boldsymbol{x}^\intercal \mathbb{E}\left[\boldsymbol{u}\right] + c^2.\end{aligned}$$

Only relevant characteristics of $\mathbb{P}$ are its

$\boldsymbol{\mu}$	$:= \mathbb{E}\left[\boldsymbol{u}\right]$	**first moment** vector
$\boldsymbol{D}$	$:= \mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}^\intercal\right]$	**second moment** matrix

Observation: gain is linear in $\boldsymbol{\mu}$ and in $\boldsymbol{D}$

Gain expansion

For randomised prediction $\boldsymbol{u} \sim \mathbb{P}$:

$$\begin{aligned}\mathbb{E}\left[\left(\boldsymbol{u}^\intercal \boldsymbol{x} + c\right)^2\right] &= \mathbb{E}\left[\boldsymbol{x}^\intercal \boldsymbol{u}\boldsymbol{u}^\intercal \boldsymbol{x} + 2c\boldsymbol{x}^\intercal \boldsymbol{u} + c^2\right] \\ &= \boldsymbol{x}^\intercal \mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}^\intercal\right]\boldsymbol{x} + 2c\boldsymbol{x}^\intercal \mathbb{E}\left[\boldsymbol{u}\right] + c^2.\end{aligned}$$

Only relevant characteristics of $\mathbb{P}$ are its

$\boldsymbol{\mu} := \mathbb{E}[\boldsymbol{u}]$	**first moment** vector
$\boldsymbol{D} := \mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}^\intercal\right]$	**second moment** matrix

Observation: gain is linear in $\boldsymbol{\mu}$ and in $\boldsymbol{D}$

Idea: forget $\mathbb{P}$ - use $\boldsymbol{\mu}$ and $\boldsymbol{D}$ as a parameter

Gain expansion

For randomised prediction $\boldsymbol{u} \sim \mathbb{P}$:

$$\mathbb{E}\left[\left(\boldsymbol{u}^\intercal \boldsymbol{x} + c\right)^2\right] = \mathbb{E}\left[\boldsymbol{x}^\intercal \boldsymbol{u}\boldsymbol{u}^\intercal \boldsymbol{x} + 2c\boldsymbol{x}^\intercal \boldsymbol{u} + c^2\right]$$
$$= \boldsymbol{x}^\intercal \mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}^\intercal\right]\boldsymbol{x} + 2c\boldsymbol{x}^\intercal \mathbb{E}\left[\boldsymbol{u}\right] + c^2.$$

Only relevant characteristics of $\mathbb{P}$ are its

$\boldsymbol{\mu} := \mathbb{E}[\boldsymbol{u}]$ **first moment** vector

$\boldsymbol{D} := \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\intercal]$ **second moment** matrix

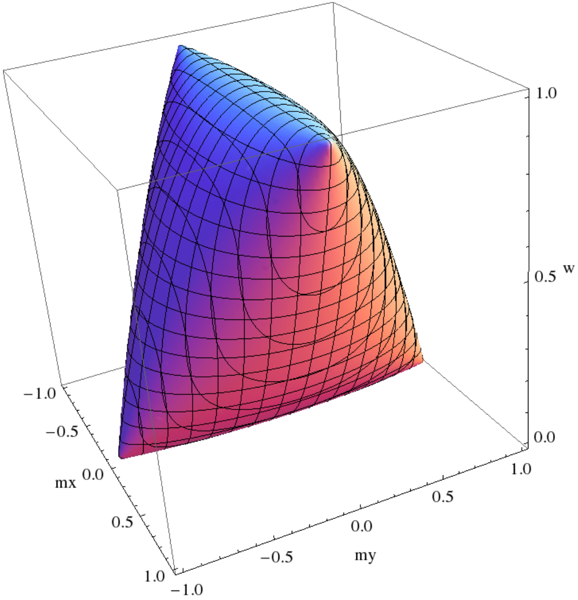
Observation: gain is linear in $\boldsymbol{\mu}$ and in $\boldsymbol{D}$

Idea: forget $\mathbb{P}$ - use $\boldsymbol{\mu}$ and $\boldsymbol{D}$ as a parameter

Careful: not all $\langle \boldsymbol{\mu}, \boldsymbol{D} \rangle$ are moments of some $\mathbb{P}$!
Parameters must lie in the **überplex**

$$\mathcal{U} := \left\{\langle \boldsymbol{\mu}, \boldsymbol{D}\rangle \;\middle|\; \exists \mathbb{P} : \boldsymbol{\mu}, \boldsymbol{D} \text{ are } 1^{\text{st}}/2^{\text{nd}} \text{ moment of } \mathbb{P}\right\}$$

Überplex $\mathcal{U}$



Characterisation

Theorem

$$\langle \boldsymbol{\mu}, \boldsymbol{D} \rangle \in \mathcal{U} \qquad \textit{iff} \qquad \mathrm{tr}(\boldsymbol{D}) = 1 \quad \textit{and} \quad \boldsymbol{D} \succeq \boldsymbol{\mu}\boldsymbol{\mu}^\intercal$$

Characterisation

Theorem

$$\langle \boldsymbol{\mu}, \boldsymbol{D} \rangle \in \mathcal{U} \qquad \textit{iff} \qquad \mathrm{tr}(\boldsymbol{D}) \;=\; 1 \quad \textit{and} \quad \boldsymbol{D} \;\succeq\; \boldsymbol{\mu}\boldsymbol{\mu}^\intercal$$

Why this is important?

- Überplex $\mathcal{U}$ is convex
- Constraint is semi-definite
- Efficient numerical linear/convex optimization over $\mathcal{U}$

Offline problem

$$\max_{(\boldsymbol{\mu},\boldsymbol{D})\in\mathcal{U}} \sum_{t=1}^{T} \left(\boldsymbol{x}_t^\intercal \boldsymbol{D} \boldsymbol{x}_t + 2c\boldsymbol{\mu}^\intercal \boldsymbol{x}_t + c^2\right)$$

Semi-definite optimisation problem
Good numerical methods

“Our” algorithm: gradient descent

Maintains the two moments $(\boldsymbol{\mu}_t, \boldsymbol{D}_t) \in \mathcal{U}$ as parameter
At trial $t = 1 \ldots T$

1. Mill *decomposes* parameter $(\boldsymbol{\mu}_t, \boldsymbol{D}_t)$ into a mixture of 6 directions and draws a direction $\boldsymbol{u}_t$ at random from it
2. Wind reveals direction $\boldsymbol{x}_t \in \mathbb{R}^2$
3. Mill receives expected gain $\mathbb{E}\left[(\boldsymbol{u}_t^\intercal \boldsymbol{x}_t + c)^2\right]$
4. Mill *updates* $(\boldsymbol{\mu}_t, \boldsymbol{D}_t)$ to $(\widehat{\boldsymbol{\mu}}_{t+1}, \widehat{\boldsymbol{D}}_{t+1})$ based on the gradient of the expected gain on $\boldsymbol{x}_t$

$$\widehat{\boldsymbol{\mu}}_{t+1} := \boldsymbol{\mu}_t + 2\eta c\, \boldsymbol{x}_t \qquad \text{and} \qquad \widehat{\boldsymbol{D}}_{t+1} := \boldsymbol{D}_t + \eta\, \boldsymbol{x}_t \boldsymbol{x}_t^\intercal$$

5. Mill produces new parameter $(\boldsymbol{\mu}_{t+1}, \boldsymbol{D}_{t+1})$ by *projecting* $(\widehat{\boldsymbol{\mu}}_{t+1}, \widehat{\boldsymbol{D}}_{t+1})$ back into the überplex

$$(\boldsymbol{\mu}_{t+1}, \boldsymbol{D}_{t+1}) := \underset{(\boldsymbol{\mu}, \boldsymbol{D}) \in \mathcal{U}}{\operatorname{argmin}} \|\boldsymbol{D} - \widehat{\boldsymbol{D}}_{t+1}\|_F^2 + \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{t+1}\|^2$$

Guarantees

regret $:=$ hindsight-optimal gain − actual gain of Mill

Theorem

The expected regret after T trials of the GD algorithm with learning rate $\eta = \sqrt{\frac{3/2}{(4c^2+1)T}}$ and initial parameters $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{D}_1 = \frac{1}{2}\boldsymbol{I}$ is upper bounded by $\sqrt{3(4c^2+1)T}$

Guarantees

regret $:=$ hindsight-optimal gain − actual gain of Mill

Theorem

The expected regret after T trials of the GD algorithm with learning rate $\eta = \sqrt{\frac{3/2}{(4c^2+1)T}}$ and initial parameters $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{D}_1 = \frac{1}{2}\boldsymbol{I}$ is upper bounded by $\sqrt{3(4c^2+1)T}$

- Regret grows sub-linearly with T
- Mill turned close to the best orientation
- Holland is saved ☺

Conclusion

- An efficient method for orienting windmills
- Characterization of set of first two moments of distributions on directions

We can do more

- Work in $n > 3$ dimensions
- Learn sets of $k \geq 1$ orthogonal directions

The hard part is to decompose a parameter $(\boldsymbol{\mu}, \boldsymbol{D})$ into a small mixture from which you can sample

Give some more details pls