Linear Pattern Recognition

Prediction Learning and Games: Chapter 11

David Lisuk

February 20, 2014

Agenda

Agenda

1. Motivation
   - Sequence Prediction with Experts
   - Sequence Prediction with Side Data
2. Mathematical Background
3. Review of Expert Prediction
4. Linear Pattern Recognition
5. Conclusion

Motivation

Typical Expert Setting

Decision Space $f_{i,t}, \hat{p}_t \in \mathcal{D}$

Outcome Space $y_t \in \mathcal{Y}$

Loss Function $\ell : \mathcal{D} \times \mathcal{Y} \mapsto \mathbb{R}$

1. Environment reveals n expert values $f_{i,t}$ for $i \in \{1, ..., n\}$
2. Forecaster make prediction $\hat{p}_t$ using expert values
3. Environment reveals truth y_t
4. Every expert and the forecaster suffer loss via loss function ℓ
5. Regret is $\max_i \sum_t \ell(\hat{p}_t, y_t) - \ell(f_{i,t}, y_t)$

Motivation

Prediction with Side Information Setting

Decision Space $\hat{p}_t \in \mathbb{R}$,

Outcome Space $y_t \in \mathbb{R}$

Loss Function $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$

Information Space $x_t \in \mathbb{R}^d$

1. Environment reveals side information x_t
2. Forecaster make prediction $\hat{p}_t = w_t \cdot x_t$ with $w_t \in \mathbb{R}^d$
3. Environment reveals truth y_t
4. Forecaster suffers loss $\ell(\hat{p}_t, y_t)$
5. Regret is $\max_u \sum_t \ell(\hat{p}_t, y_t) - \ell(u \cdot x_t, y_t)$

Each possible weight vector $w \in \mathbb{R}^d$ is an "expert"

Agenda

Legendre Functions

- A function $F : \mathcal{A} \mapsto \mathbb{R}$ is Legendre if 3 properties hold:
  1. $\mathcal{A} \subseteq \mathbb{R}^d$ is nonempty and $\mathrm{int}(\mathcal{A})$ is convex
  2. F is strictly convex and is continuously differentiable
  3. As x approaches a boundary of $\mathcal{A}$, $||\nabla F(x)|| \to \infty$
- For all Legendre functions, F there is a dual $F^\star : \mathcal{A}^\star \mapsto \mathbb{R}$
  - Defined as: $F^\star(u) = \sup_{v \in \mathcal{A}} (u \cdot v - F(v))$
  - $\mathcal{A}^\star$ is the range of $\nabla F : \mathrm{int}(\mathcal{A}) \mapsto \mathbb{R}^d$
  - $(F^\star)^\star = F$
  - Lemma 11.5: $\nabla F^\star = (\nabla F)^{-1}$

Legendre Function: Example

- The squared p-norm ($\frac{1}{2}||u||_p^2$, $p \geq 2$) is Legendre
  - $\mathcal{A} = \mathbb{R}^d$, (obviously non empty and convex)
  - All norm functions are convex
  - $(\nabla F(x))_i = \frac{\text{sign}(x_i)|x_i|^{p-1}}{||x||_p^{p-2}}$ which goes to ∞ as x_i does
- $F^\star = \frac{1}{2}||u||_q^2$ such that $\frac{1}{p} + \frac{1}{q} = 1$
  - $(\nabla F(x))^{-1} = \nabla \frac{1}{2}||x||_q^2$

Bregman Divergence

- A Bregman divergence is a way of defining a distance measure using a Legendre function

Bregman Divergence on F

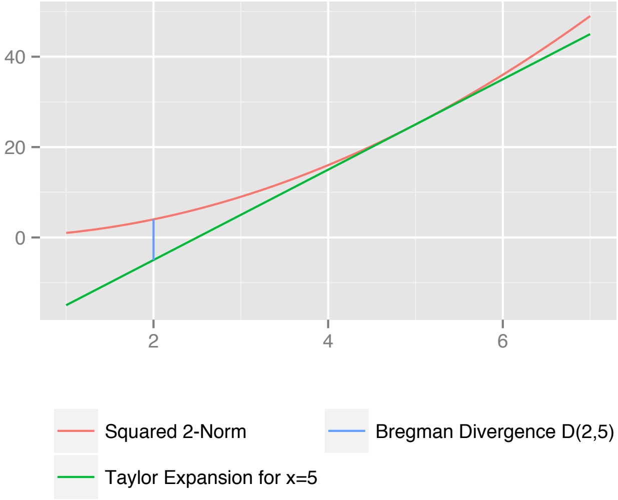
Let F be a Legendre function, then the Bregman divergence induced by F is:

$$D_F(u, v) = F(u) - F(v) - (u - v)\nabla F(v)$$

- The difference between $F(u)$ and its first order Taylor approximation about v
- Lemma 11.1:

$$D_F(u, v) + D_F(v, w) = D_F(u, w) + (u - v)(\nabla F(w) - \nabla F(v))$$

Bregman Divergence: Visual Intuition



Bregman Projections

Bregman Projection

A Bregman projection of v onto a convex set S is defined as:

$$\mathcal{P}_F(v) = \text{argmin}_{u \in S} D_F(u, v)$$

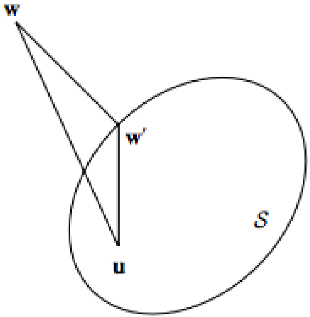
This is the point in S closest to v, as defined by D_F

Generalized Pythagorean Inequality

Generalized Pythagorean Inequality

For all $w \in \text{int}(\mathcal{A})$, and all convex and closed sets $S \subseteq \mathbb{R}^d$ with $S \cap \mathcal{A} \neq \emptyset$, and $w' = \mathcal{P}_F(w)$:

$$D_F(u, w) \geq D_F(u, w') + D_F(w', w) \; \forall u \in S$$



Proof of Generalized Pythagorean Inequality

- Define $G(x) = D(x, w) - D(x, w')$, expanding shows this is linear
- Let $x_\alpha = \alpha u + (1 - \alpha) w'$
- Due to linearity we get: $G(x_\alpha) = \alpha G(u) + (1 - \alpha) G(w')$
- Expanding:
  $D(x_\alpha, w) - D(x_\alpha, w') = \alpha(D(u, w) - D(u, w')) + (1 - \alpha) D(w', w)$
- Rearranging and assuming $\alpha > 0$:

$$\frac{D(x_\alpha, w) - D(x_\alpha, w') - D(w', w)}{\alpha} = D(u, w) - D(u, w') - D(w', w)$$

- Since w' was chosen to be point closest to w in S, $D(x_\alpha, w) \geq D(w', w)$ thus:

$$\frac{D(x_\alpha, w) - D(x_\alpha, w') - D(w', w)}{\alpha} \geq \frac{-D(x_\alpha, w')}{\alpha}$$

Proof of Generalized Pythagorean Inequality

- Rearranging gives:

$$D(u, w) + \frac{-D(x_\alpha, w')}{\alpha} \geq D(u, w') + D(w', w)$$

- Thus we must show an $\alpha > 0$ exists st:

$$\frac{-D(x_\alpha, w')}{\alpha} = 0$$

- At the limit this is true:

$$\lim_{\alpha \to 0^+} \frac{-D(x_\alpha, w')}{\alpha} = \lim_{\alpha \to 0^+} \frac{-D(w' + \alpha(u - w'), w') - D(w')}{\alpha}$$

- The rhs is the derivative of D at w' in the direction $u - w'$
- Since $D(w', w') = 0$, and D is non-negative, D' must be 0
- Hence the Generalized Pythagorean Inequality is true

Agenda

Weighted Average Predictor

- Predict the weighted average of the expert advice:

$$\hat{p}_t = \frac{\sum_{i=1}^{N} w_{i,t-1} f_{i,t}}{\sum_{j=1}^{N} w_{j,t-1}}$$

- Define w as the derivative of a *potential function*(Φ) of regret.

$$\Phi(u) = \psi\left(\sum_{i=1}^{N} \phi(u_i)\right)$$

  - $\phi : \mathbb{R} \mapsto \mathbb{R}$ is non-negative, increasing, and twice differentiable
  - $\psi : \mathbb{R} \mapsto \mathbb{R}$ is non-negative, strictly increasing, concave, and twice differentiable

- Define weights with this potential function:

$$w_{t-1} = \nabla\Phi(R_{t-1})$$

Weighted Average Predictor

- The "Blackwell Condition" states:

$$\sup_{y_t \in \mathcal{Y}} r_t \cdot \nabla \Phi(R_{t-1}) \leq 0$$

- The potential gradient and instantaneous regret point away from each other
- Thus the potential stays near its minimum
- Theorem 2.1:

$$\Phi(R_n) \leq \Phi(0) + \frac{1}{2} \sum_{t=1}^{n} C(r_t)$$

$$C(r_t) = \sup_{u \in \mathbb{R}^N} \psi' \left(\sum_{i=1}^{N} \phi(u_i) \right) \sum_{i=1}^{N} \phi''(u_i) r_{i,t}^2$$

Exponentially Weighted Average Forecaster

- Setting the potential to be:

$$\Phi_\eta(u) = \frac{1}{\eta} \log \sum_{i=1}^{N} \exp(\eta u_i)$$

- Plugging this potential into theorem 2.1 leads to this regret bound:

$$\max_i R_{i,n} \leq \frac{\log N}{\eta} + \frac{n\eta}{2}$$

- Tighter bounds are possible with specific loss functions

Agenda

Linear Pattern Recognition as Experts

The linear pattern recognition problem is very similar to experts, can we use the same algorithm?

NO

- In experts we measure regret vs best expert (finite number of experts)
- In linear pattern recognition we compare to best weight vector (infinite set)

However, using potentials and Legendre dual we can come up with a modified algorithm.

Potential-Based Gradient Descent

- Choose a potential Φ meeting previous requirements *AND* that is *Legendre*
- Then define $w_{t-1} = \nabla\Phi(R_{t-1})$
- Since Φ is Legendre, we get the following:

$$R_{t-1} = \nabla\Phi^{\star}(w_t)$$

Key Idea

Since we can't easily search for the w_t which did the best in the past, this dual formulation allows us to directly minimize our increase in regret.

Regret Updating

- Define $\theta_t = R_t$ to reinforce the notion that regret is minimized

Regret Update Rules

Primal	*Dual*
$\theta_t = \theta_{t-1} + r_t$	$\nabla\Phi^\star(w_t) = \nabla\Phi^\star(w_{t-1}) + r_t$

- After updating regret, we then use duality to update weights

$$w_t = \nabla\Phi(\theta_t)$$

Potential-Based Gradient Descent:Visual Intuition

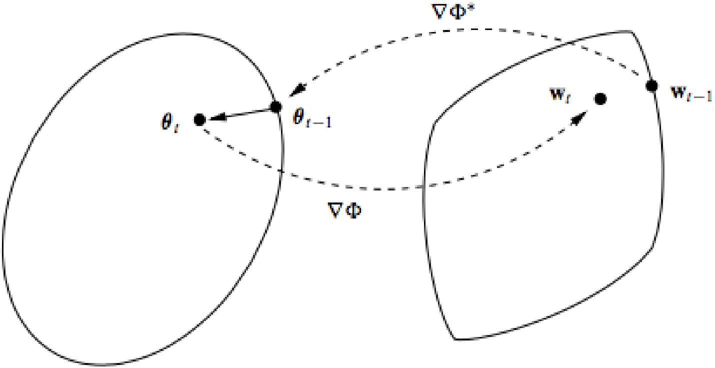


Figure: Duality of regret(left) and weight(right) space

Weight Updating

- This duality argument implies the weight update rule:

$$w_t = \nabla\Phi\left(\nabla\Phi^\star(w_{t-1}) + r_t\right)$$

- However, r_t still depends on the optimal weight
- r_t can be approximated by the *loss gradient*

Final Update Rule

$$w_t = \nabla\Phi\left(\nabla\Phi^\star(w_{t-1}) - \lambda\nabla\ell(x_t \cdot w_{t-1}, y_t)\right)$$

With $\lambda \geq 0$ being an arbitrary scale factor

Final Algorithm

1. Receive x_t from environment
2. Make prediction $\hat{p}_t = w_{t-1} \cdot x_t$
3. Receive y_t from environment
4. Incur loss $\ell(w_{t-1}) = \ell_t(\hat{p}_t, y_t)$
5. Update weights $w_t = \nabla\Phi\left(\nabla\Phi^\star(w_{t-1}) - \lambda\nabla\ell(x_t \cdot w_{t-1}, y_t)\right)$

Bound for Arbitrary Potential

Theorem 11.1

$$R_n(u) \le \frac{1}{\lambda} D_{\Phi^\star}(u, w_0) + \frac{1}{\lambda} \sum_{t=1}^{n} D_{\Phi^\star}(w_{t-1}, w_t)$$

Proof:

$$\begin{aligned}
\ell_t(w_{t-1}) &\le \ell_t(u) - (u - w_{t-1}) \cdot \nabla \ell_t(w_{t-1}) \\
&= \ell_t(u) + \frac{1}{\lambda}(u - w_{t-1}) \cdot (\nabla \Phi^\star(w_t) - \nabla \Phi^\star(w_{t-1})) \\
&= \ell_t(u) + \frac{1}{\lambda}(D_{\Phi^\star}(u, w_{t-1}) - D_{\Phi^\star}(u, w_t) + D_{\Phi^\star}(w_{t-1}, w_t))
\end{aligned}$$

We then sum over t and drop $-D_{\Phi^\star}(u, w_n)$ to complete to proof

Agenda

Conclusion

- This covered the basics of Linear Pattern Recognition
- Much more in the book:
  - Using transfer functions
  - Tracking weight vectors
  - Time varying potentials
  - etc
- Also chapter 12 extends this to linear classification

Useful Resources

Prediction Learning and Games Nicolò Cesa-Bianchi and Gábor Lugosi

Bregman Divergence http://mark.reid.name/blog/meet-the-bregman-divergences.html