Lecture 11 and 12: Probabilistic Ranking

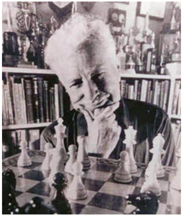
4F13: Machine Learning

Joaquin Quiñonero-Candela and Carl Edward Rasmussen

Department of Engineering
University of Cambridge

http://mlg.eng.cam.ac.uk/teaching/4f13/

Motivation for ranking



Competition is central to our lives. It is an innate biological trait, and the driving principle of many sports.

In the Ancient Olympics (700 BC), the winners of the events were admired and immortalised in poems and statues.

Today in pretty much every sport there are player or team rankings. (Football leagues, Poker tournament rankings, etc).

We are going to focus on one example: *tennis players*, in men singles games.

We are going to keep in mind the goal of answering the following question: *What is the probability that player 1 defeats player 2?*

The ATP ranking system for tennis players

Men Singles ranking as of 28 December 2011:

Rank, Name & Nationality	Points	Tournaments Played
1 Djokovic, Novak (SRB)	13,675	19
2 Nadal, Rafael (ESP)	9,575	20
3 Federer, Roger (SUI)	8,170	19
4 Murray, Andy (GBR)	7,380	19
5 Ferrer, David (ESP)	4,880	23
6 Tsonga, Jo-Wilfried (FRA)	4,335	25
7 Berdych, Tomas (CZE)	3,700	24
8 Fish, Mardy (USA)	2,965	24
9 Tipsarevic, Janko (SRB)	2,595	28
10 Almagro, Nicolas (ESP)	2,380	27
11 Del Potro, Juan Martin (ARG)	2,315	22
12 Simon, Gilles (FRA)	2,165	28
13 Soderling, Robin (SWE)	2,120	22
14 Roddick, Andy (USA)	1,940	20
15 Monfils, Gael (FRA)	1,935	23
16 Dolgopolov, Alexandr (UKR)	1,925	30
17 Wawrinka, Stanislas (SUI)	1,820	23
18 Isner, John (USA)	1,800	25
19 Gasquet, Richard (FRA)	1,765	21
20 Lopez, Feliciano (ESP)	1,755	28

ATP: Association of Tennis Professionals (`www.atpworldtour.com`)

- Sum of points from best 18 results of the past 52 weeks.
- Mandatory events: 4 Grand Slams, and 8 Masters 1000 Series events.
- Best 6 results from International Events (4 of these must be 500 events).

Points breakdown for all tournament categories (2012):

	W	F	SF	QF	R16	R32	R64	R128	Q
Grand Slams	2000	1200	720	360	180	90	45	10	25
Barclays ATP World Tour Finals	*1500								
ATP World Tour Masters 1000	1000	600	360	180	90	45	10(25)	(10)	(1)25
ATP 500	500	300	180	90	45	(20)			(2)20
ATP 250	250	150	90	45	20	(5)			(3)12
Challenger 125,000 +H	125	75	45	25	10				5
Challenger 125,000	110	65	40	20	9				5
Challenger 100,000	100	60	35	18	8				5
Challenger 75,000	90	55	33	17	8				5
Challenger 50,000	80	48	29	15	7				3
Challenger 35,000 +H	80	48	29	15	6				3
Futures** 15,000 +H	35	20	10	4	1				
Futures** 15,000	27	15	8	3	1				
Futures** 10,000	18	10	6	2	1				

The Grand Slams are the Australian Open, the French Open, Wimbledon, and the US Open.
The Masters 1000 Tournaments are: Cincinnati, Indian Wells, Madrid, Miami, Monte-Carlo, Paris, Rome, Shanghai, and Toronto.
The Masters 500 Tournaments are: Acapulco, Barcelona, Basel, Beijing, Dubai, Hamburg, Memphis, Rotterdam, Tokyo, Valencia and Washington.
The Masters 250 Tournaments are: Atlanta, Auckland, Bangkok, Bastad, Belgrade, Brisbane, Bucharest, Buenos Aires, Casablanca, Chennai, Delray Beach, Doha, Eastbourne, Estoril, Gstaad, Halle, Houston, Kitzbuhel, Kuala Lumpur, London, Los Angeles, Marseille, Metz, Montpellier, Moscow, Munich, Newport, Nice, Sao Paulo, San Jose, ’s-Hertogenbosch, St. Petersburg, Stockholm, Stuttgart, Sydney, Umag, Vienna, Vina del Mar, Winston-Salem, Zagreb and Dusseldorf.

A laundry list of objections and open questions

Rank, Name & Nationality	Points
1 Djokovic, Novak (SRB)	13,675
2 Nadal, Rafael (ESP)	9,575
3 Federer, Roger (SUI)	8,170
4 Murray, Andy (GBR)	7,380

Some questions:

- Is a player ranked higher than another more likely to win?
- What is the probability that Murray defeats Djokovic?
- *How much would you (rationally) bet on Murray?*

And some concerns:

- The points system ignores who you played against.
- 6 out of the 18 tournaments don't need to be common to two players.

Other examples: Premier League. Meaningless intermediate results throughout the season: doesn't say whom you played and whom you didn't!

Towards a probabilistic ranking system

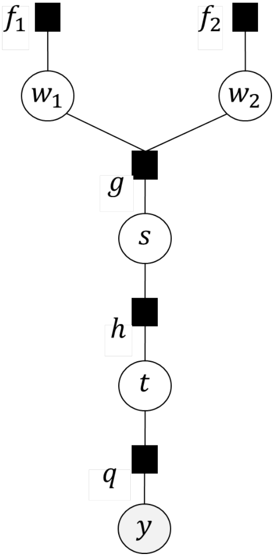
What we really want is to infer is a player's *skill*.

- Skills must be comparable: a player of higher skill is more likely to win.
- We want to do probabilistic inference of players skills.
- We want to be able to compute the probability of a game outcome.

A generative model for game outcomes:

1. Take two tennis players with known *skills* ($w_i \in \mathbb{R}$)
   - Player 1 with skill w_1.
   - Player 2 with skill w_2.
2. Compute the difference between the skills of Player 1 and Player 2:
   $$s = w_1 - w_2$$
3. Add noise ($n \sim \mathcal{N}(0, 1)$) to account for *performance* inconsistency:
   $$t = s + n$$
4. The game outcome is given by $y = \text{sign}(t)$
   - $y = +1$ means Player 1 wins.
   - $y = -1$ means Player 2 wins.

TrueSkill™, a probabilistic skill rating system



- w_1 and w_2 are the skills of Players 1 and 2. We treat them in a Bayesian way:

$$\text{prior} \qquad p(w_i) = \mathcal{N}(w_i; \mu_i, \sigma_i^2)$$

- $s = w_1 - w_2$ is the *skill difference*.
- $t \sim \mathcal{N}(t; s, 1)$ is the *performance difference*.
- The probability of outcome y given known skills is:

$$p(y|w_1, w_2) = \iint p(y|t)p(t|s)p(s|w_1, w_2)dsdt$$

likelihood

- The posterior over skills given the game outcome is:

$$p(w_1, w_2|y) = \frac{p(w_1)p(w_2)p(y|w_1, w_2)}{\iint p(w_1)p(w_2)p(y|w_1, w_2)dw_1dw_2}$$

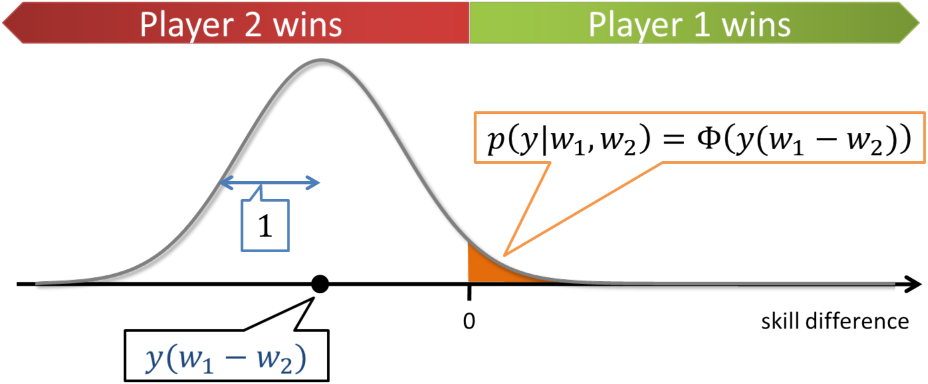
TrueSkill™: A Bayesian Skill Rating System. Herbrich, Minka and Graepel, NIPS19, 2007.

The likelihood in detail

$$
\begin{aligned}
p(y|w_1, w_2) &= \iint p(y|t)p(t|s)p(s|w_1, w_2)\,ds\,dt = \int p(y|t)p(t|w_1, w_2)\,dt \\
&= \int_{-\infty}^{+\infty} \delta(y - \mathrm{sign}(t))\mathcal{N}(t; w_1 - w_2, 1)\,dt \\
&= \int_{-\infty}^{+\infty} \delta(1 - \mathrm{sign}(yt))\mathcal{N}(yt; y(w_1 - w_2), 1)\,dt \\
&= y \int_{-y\infty}^{+y\infty} \delta(1 - \mathrm{sign}(z))\mathcal{N}(z; y(w_1 - w_2), 1)\,dz \qquad (\text{use } \ z \equiv yt) \\
&= \int_{-\infty}^{+\infty} \delta(1 - \mathrm{sign}(z))\mathcal{N}(z; y(w_1 - w_2), 1)\,dz \\
&= \int_{0}^{+\infty} \mathcal{N}(z; y(w_1 - w_2), 1)\,dz = \int_{-\infty}^{y(w_1 - w_2)} \mathcal{N}(x; 0, 1)\,dx \qquad (\text{use } x \equiv y(w_1 - w_2) - z) \\
&= \Phi(y(w_1 - w_2)) \qquad \Big(\text{where } \Phi(a) = \int_{-\infty}^{a} \mathcal{N}(x; 0, 1)\,dx\Big)
\end{aligned}
$$

$\Phi(a)$ is the Gaussian cumulative distribution function, or 'probit' function.

The likelihood in a picture



An intractable posterior

The joint posterior distribution over skills does not have a closed form:

$$p(w_1, w_2|y) = \frac{\mathcal{N}(w_1; \mu_1, \sigma_1^2)\mathcal{N}(w_2; \mu_2, \sigma_2^2)\Phi(y(w_1 - w_2))}{\iint \mathcal{N}(w_1; \mu_1, \sigma_1^2)\mathcal{N}(w_2; \mu_2, \sigma_2^2)\Phi(y(w_1 - w_2))dw_1dw_2}$$

- w_1 and w_2 become correlated, the posterior does not factorise.
- The posterior is no longer a Gaussian density function.

The normalising constant of the posterior, the prior over y does have closed form:

$$p(y) = \iint \mathcal{N}(w_1; \mu_1, \sigma_1^2)\mathcal{N}(w_2; \mu_2, \sigma_2^2)\Phi(y(w_1 - w_2))dw_1dw_2 = \Phi\Big(\frac{y(\mu_1 - \mu_2)}{\sqrt{1 + \sigma_1^2 + \sigma_2^2}}\Big)$$

This is a smoother version of the likelihood $p(y|w_1, w_2)$.

Can you explain why?

Gibbs sampling for the TrueSkill model

We have $g = 1, \ldots, N$ games where I_g: id of Player 1 and J_g: id of Player 2. The outcome of game g is $y_g = +1$ if I_g wins, $y_g = -1$ if J_g wins.

Gibbs sampling alternates between sampling skills $\mathbf{w} = [w_1, \ldots, w_M]^\top$ conditional on fixed performance differences $\mathbf{t} = [t_1, \ldots, t_N]^\top$, and sampling $\mathbf{t}$ conditional on fixed $\mathbf{w}$.

1. Initialise $\mathbf{w}$, e.g. from the prior $p(\mathbf{w})$.
2. Sample the *performance differences* from
$$p(t_g | w_{I_g}, w_{J_g}, y_g) \;\propto\; \delta(y_g - \mathrm{sign}(t_g))\mathcal{N}(t_g; w_{I_g} - w_{J_g}, 1)$$
3. Jointly sample the *skills* from
$$\underbrace{p(\mathbf{w}|\mathbf{t})}_{\mathcal{N}(\mathbf{w};\mu,\Sigma)} \propto \underbrace{p(\mathbf{w})}_{\mathcal{N}(\mathbf{w};\mu_0,\Sigma_0)} \prod_{g=1}^{N} \underbrace{p(t_g | w_{I_g}, w_{J_g})}_{\propto \mathcal{N}(\mathbf{w};\mu_g,\Sigma_g)}$$
4. Go back to step 2.

Gaussian identities

The conditional for the performance is both Gaussian in t_g and proportional to a Gaussian in $\mathbf{w}$

$$\begin{aligned} p(t_g|w_{I_g}, w_{J_g}) &\propto \exp\big(-\tfrac{1}{2}(w_{I_g} - w_{J_g} - t_g)^2\big) \\ &\propto \mathcal{N}\Big(-\tfrac{1}{2}\left(\begin{array}{c} w_{I_g} - \mu_1 \\ w_{J_g} - \mu_2 \end{array}\right)^\top \left[\begin{array}{rr} 1 & -1 \\ -1 & 1 \end{array}\right] \left(\begin{array}{c} w_{I_g} - \mu_1 \\ w_{J_g} - \mu_2 \end{array}\right)\Big) \end{aligned}$$

with $\mu_1 - \mu_2 = t_g$. Notice that

$$\left[\begin{array}{rr} 1 & -1 \\ -1 & 1 \end{array}\right] \left(\begin{array}{c} \mu_1 \\ \mu_2 \end{array}\right) = \left(\begin{array}{r} t_g \\ -t_g \end{array}\right)$$

Remember that for products of Gaussians precisions add up, and means weighted by precisions (natural parameters) also add up:

$$\mathcal{N}(\mathbf{w}; \mu_a, \Sigma_a)\mathcal{N}(\mathbf{w}; \mu_b, \Sigma_b) = z_c \mathcal{N}\big(\mathbf{w}; \Sigma_c(\Sigma_a^{-1}\mu_a + \Sigma_b^{-1}\mu_b), \Sigma_c = (\Sigma_a^{-1} + \Sigma_b^{-1})^{-1}\big)$$

Conditional posterior over skills given performances

We can now compute the covariance and the mean of the conditional posterior.

$$\Sigma^{-1} = \Sigma_0^{-1} + \underbrace{\sum_{g=1}^{N} \Sigma_g^{-1}}_{\tilde{\Sigma}^{-1}} \qquad\qquad \mu = \Sigma\big(\Sigma_0^{-1}\mu_0 + \underbrace{\sum_{g=1}^{N} \Sigma_g^{-1}\mu_g}_{\tilde{\mu}}\big)$$

To compute the mean it is useful to note that:

$$\tilde{\mu}_i = \sum_{g=1}^{N} t_g\big(\delta(i - I_g) - \delta(i - J_g)\big)$$

And for the covariance we note that:

$$[\tilde{\Sigma}^{-1}]_{ii} = \sum_{g=1}^{N} \delta(i - I_g) + \delta(i - J_g)$$

$$[\tilde{\Sigma}^{-1}]_{i\neq j} = -\sum_{g=1}^{N} \delta(i - I_g)\delta(j - J_g) + \delta(i - J_g)\delta(j - I_g)$$