Regression-Based Approaches

Logistic Regression: Xi(Q,D) is feature; β's are parameters

$$\log \frac{P(R=1|Q,D)}{1-P(R=1|Q,D)} = \beta_0 + \sum_{i=1}^{n} \beta_i X_i$$

Estimate β's by maximizing the likelihood of training data

$$P(R=1|Q,D) = \frac{1}{1+\exp(-\beta_0 - \sum_{i=1}^{n} \beta_i X_i)}$$

	X1(Q,D) BM25	X2 (Q,D) PageRank	X3(Q,D) BM25Anchor
D1 (R=1)	0.7	0.11	0.65
D2 (R=0)	0.3	0.05	0.4

$$p(\{(Q,D_1,1),(Q,D_2,0)\}) = \frac{1}{1+\exp(-\beta_0 - 0.7\beta_1 - 0.11\beta_2 - 0.65\beta_3)} * (1 - \frac{1}{1+\exp(-\beta_0 - 0.3\beta_1 - 0.05\beta_2 - 0.4\beta_3)})$$

$$\vec{\beta}^* = \arg\max_{\vec{\beta}} p(\{(Q_1,D_{11},R_{11}),(Q_1,D_{12},R_{12}),....,(Q_n,D_{m1},R_{m1}),...\})$$

Once β's are known, we can take Xi(Q,D) computed based on a new query and a new document to generate a score for D w.r.t. Q.