CS 4501: INFORMATION RETRIEVAL

HW2— Basic Concepts about Probability and Linear Algebra

This homework assignment is designed to help you recall important concepts in probability and linear algebra. This assignment has in total **120** points, among which 20 points for the last bonus question. The deadline and submission guidance are explicitly described here.

1. Joint, marginal and conditional probabilities (35pts)

	$Y = 1$	$Y = 2$
$X = 1$	1/3	1/12
$X = 2$	1/6	0
$X = 4$	1/12	1/3

1.1 Compute $P(X \leq 2, Y > 1)$. (5pts)
1.2 Compute marginal probability mass function for X and Y. (5pts)
1.3 Compute $P(Y = 2 \mid X = 1)$. (5pts)
1.4 Are X and Y independent? (5pts)
1.5 Define $Z = X - 2Y$, compute $P(X = 2 \mid Z = 0)$. (5pts)

1.6 Compute $E[X|Y=1]$. (5pts)

1.7 Compute $Var[X|Y=2]$. (5pts)

2. Proof of probabilistic ranking principle (15pts)

In our lecture discussion, we introduce the concept of probabilistic ranking principle from risk minimization perspective. However, we have not yet proved that ranking by the probability of being relevant will minimize the risk.
Please finish the proof by assuming: 1) probability of a document being relevant is independent of the other documents; 2) users will sequentially browsing the results from top to bottom. Hint: what you need to prove is that whenever a user stops browsing, the scanned results are optimal in terms of those two types of risk, i.e., presenting an irrelevant result or missing a relevant result.

3. Maximum likelihood estimation (20pts)

The probability density function of Gaussian distribution is $f(x|\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Given a set of observations from this Gaussian distribution, $X=x_1,x_2,\ldots,x_n$, estimate the mean parameter μ and variance parameter σ with details.

4. Cosine v.s., Euclidean distance (15pts)

x and y are two n-dimensional unit vectors, i.e., $\sum_{i=1}^{n}x_i^2=1$. Figure out the relationship between the cosine similarity between x and y and Euclidean distance between x and y. *Hint: can you compute cosine similarity from Euclidean distance, and vice versus.*

5. α_d in language model ranking modules (15pts)

The generic ranking function based on smoothed language models can be written as,

$$\log P(q \mid d) = \sum_{w \in d \cap q} \left(\log \frac{p(w \mid d)}{\alpha_d p(w \mid C)} \right) + |q| \log \alpha_d$$

where $p(w \mid d)$ is the smoothed language model for document d, $p(w \mid C)$ is a reference language model, and α_d is a smoothing parameter defined in our lecture slide for "Refine the idea of smoothing."

Please derive the value of α_d for Dirichlet prior smoothing parameterized by μ. *Hint: build the connection between α_d and Dirichlet prior smoothing parameter μ.*

6. Bonus question: α_d in language model ranking modules (20pts)

Using the same analysis technique you have developed in Problem 5 and derive the corresponding α_d for additive smoothing (i.e., adding δ) and linear interpolation smoothing (parameterized by λ). 10 points for each result.

Deadline and How to submit

The deadline for this assignment is 11:55pm, Monday, October 19th.

Please submit a **PDF** version of your solutions to our Collab site before the

deadline. Your grade will be announced in Collab after the deadline.

Share on Twitter Share on Facebook Share on Google+

Updated October 11, 2015

Department of Computer Science University of Virginia

© 2015 CS 4501: Information Retrieval powered by Jekyll + Skinny Bones.