CS613 \ Assignment #1 Solutions

Ariel Stolerman

1)

Let $x = \{x_n | \, n = 1, \dots, N\}$ denote a set of data points that are i.i.d. Gaussian.

From that we know that:

$$p(x|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) = \boxed{\prod_{n=1}^{N} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2}}$$

And so $\ln p(x|\mu, \sigma^2)$ is:

$$\ln \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}(x_n|\mu, \sigma^2) = \sum_{n=1}^{N} \ln \left(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2} \right) = \sum_{n=1}^{N} \left[\ln(2\pi\sigma^2)^{-\frac{1}{2}} - \frac{1}{2\sigma^2}(x_n - \mu)^2 \right] =$$

$$\boxed{-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi)}$$

1.1)

Proof for $\boldsymbol{\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n}$:

We take first derivative of $\ln p(x|\mu, \sigma^2)$ w.r.t. μ:

$$\frac{\partial(\ln p(x|\mu, \sigma^2))}{\partial \mu} = \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu) = 0 \Leftrightarrow \sum_{n=1}^{N} (x_n - \mu) = 0 \Leftrightarrow \sum_{n=1}^{N} x_n = \sum_{n=1}^{N} \mu = N\mu \Rightarrow \boxed{\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n}$$

We take second derivative to check it is a maximum:

$\frac{\partial \left(\frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu) \right)}{\partial \mu} = -\frac{N}{\sigma^2} < 0 \Rightarrow$ it is a maximum.

Proof for $\boldsymbol{\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2}$:

We do the same as before w.r.t. σ:

$$\frac{\partial(\ln p(x|\mu, \sigma^2))}{\partial \sigma} = \frac{1}{\sigma^3} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{\sigma} = 0 \Leftrightarrow \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 = N \Rightarrow \boxed{\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2}$$

And making sure it is a maximum:

$$\frac{\partial \left(\frac{1}{\sigma^3} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{\sigma} \right)}{\partial \sigma} = -\frac{3}{\sigma^4} \sum_{n=1}^{N} (x_n - \mu)^2 + \frac{N}{\sigma^2} < 0 \Leftrightarrow \sigma^2 < 3 \cdot \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2$$

Which we know is true for $\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2$.

1.2)

We show that σ_{ML}^2 is biased by showing that $E[\sigma_{ML}^2] \neq \sigma^2$:

$$E[\sigma_{ML}^2] = E\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2\right] = E\left[\frac{1}{N}\sum_{n=1}^{N}x_n^2 - \frac{2}{N}\sum_{n=1}^{N}x_n\mu_{ML} + \frac{1}{N}\sum_{n=1}^{N}\mu_{ML}^2\right] =$$

$$\frac{1}{N}\sum_{n=1}^{N}\underbrace{E[x_n^2]}_{\mu^2+\sigma^2} - 2E\left[\mu_{ML}\cdot\underbrace{\frac{1}{N}\sum_{n=1}^{N}x_n}_{\mu_{ML}}\right] + E\left[\underbrace{\frac{1}{N}\sum_{n=1}^{N}\mu_{ML}^2}_{\mu_{ML}^2}\right] = \frac{1}{N}\sum_{n=1}^{N}(\mu^2+\sigma^2) - E\left[\underbrace{\mu_{ML}^2}_{\left(\frac{1}{N}\sum_{n=1}^{N}x_n\right)^2}\right] =$$

$$\mu^2+\sigma^2 \quad \underbrace{-\frac{1}{N^2}E\left[\left(\sum_{n=1}^{N}x_n\right)^2\right]}_{(*)\ Var[\sum_{n=1}^{N}x_n]+E[\sum_{n=1}^{N}x_n]^2} \overset{(**)}{=} \mu^2+\sigma^2-\frac{1}{N^2}(N\sigma^2+N^2\mu^2) = \frac{N^2\mu^2+N^2\sigma^2-N\sigma^2-N^2\mu^2}{N^2} = \boxed{\frac{N-1}{N}\sigma^2} \neq \sigma^2$$

$(*)\ Var[X] = E[X^2] - E[X]^2 \Rightarrow E[X^2] = Var[X] + E[X]^2$

$(**)\ Var[\sum_{n=1}^{N}x_n] = \sum_{n=1}^{N}Var[x_n] = N\sigma^2$ because the data points are i.i.d., therefore $\forall i \neq j: Cov(x_i, x_j) = 0$

Therefore σ_{ML}^2 is biased.

1.3)

Now we will use the true mean μ instead of μ_{ML} and show that σ_{ML}^2 becomes unbiased:

$$E[\sigma_{ML}^2] = E\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2\right] = E\left[\frac{1}{N}\sum_{n=1}^{N}x_n^2 - \frac{2}{N}\sum_{n=1}^{N}x_n\mu + \frac{1}{N}\sum_{n=1}^{N}\mu^2\right] =$$

$$\frac{1}{N}\sum_{n=1}^{N}\underbrace{E[x_n^2]}_{\mu^2+\sigma^2} - 2E\left[\frac{1}{N}\sum_{n=1}^{N}x_n\mu\right] + E\left[\underbrace{\frac{1}{N}\sum_{n=1}^{N}\mu^2}_{\mu^2}\right] = \mu^2+\sigma^2-2\mu\frac{1}{N}\sum_{n=1}^{N}E[x_n] + E[\mu^2] = \mu^2+\sigma^2-2\mu^2+\mu^2 = \boxed{\sigma^2}$$

Therefore σ_{ML}^2 becomes unbiased.

1.4)

Following is a derivation of the posterior of the mean $p(\mu|x)$ when using Gaussian prior $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$:

The likelihood:

$$p(x|\mu) = \prod_{n=1}^{N}p(x_n|\mu) = \prod_{n=1}^{N}\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2} = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2}$$

The prior is given:

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2) = \frac{1}{(2\pi\sigma_0^2)^{\frac{1}{2}}}e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}$$

Now the posterior:

$$p(\mu|x) \propto p(x|\mu)p(\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2} \cdot \frac{1}{(2\pi\sigma_0^2)^{\frac{1}{2}}} e^{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2}$$

Looking at the exponent:

$$-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2 = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}x_n^2 + \frac{1}{\sigma^2}\underbrace{\sum_{n=1}^{N}x_n}_{=\mu N\mu_{ML}}\mu - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\mu^2 - \frac{\mu^2}{2\sigma_0^2} + \frac{\mu\mu_0}{\sigma_0^2} - \frac{\mu_0^2}{2\sigma_0^2} =$$

$$\mu^2\left[-\frac{N}{2\sigma^2} - \frac{1}{2\sigma_0^2}\right] + \mu\left[\frac{N\mu_{ML}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right] + \left[-\frac{\sum_{n=1}^{N}x_n^2}{2\sigma^2} - \frac{\mu_0^2}{2\sigma_0^2}\right]$$

The product of two Gaussians is Gaussian itself, so this should be of the form:

$$-\frac{1}{2\sigma_N^2}(\mu-\mu_N)^2 = -\frac{1}{2\sigma_N^2}(\mu^2 - 2\mu\mu_N + \mu_N^2)$$

And now we extract σ_N^2, μ_N by comparing coefficients of the quadratic expressions:

$$-\frac{1}{2\sigma_N^2} = -\frac{N}{2\sigma^2} - \frac{1}{2\sigma_0^2} \Leftrightarrow \frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} = \frac{\sigma_0^2 N + \sigma^2}{\sigma^2\sigma_0^2} \Rightarrow \boxed{\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{\sigma_0^2 N + \sigma^2}}$$

$$\frac{\mu_N}{\sigma_N^2} = \frac{N\mu_{ML}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \Leftrightarrow \mu_N = \frac{\sigma^2\sigma_0^2}{\sigma_0^2 N + \sigma^2}\left(\frac{N\mu_{ML}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) \Rightarrow \boxed{\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0}$$

$\Rightarrow$

$$\boxed{p(\mu|x) = \mathcal{N}\left(\mu \mid \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0, \frac{\sigma^2\sigma_0^2}{\sigma_0^2 N + \sigma^2}\right)}$$

2)

Let x, y be two random variables with joint distribution $p(x, y)$. Then:

Proof for $\boldsymbol{E[x] = E_y[E_x[x|y]]}$:

$$E_y[E_x[x|y]] = E_y\left[\sum_x x \cdot p(x|y)\right] = \sum_x x \underbrace{E_y[p(x|y)]}_{(*)} = \sum_x x \cdot p(x) = \boxed{E[x]}$$

$$(*)\ E_y[p(x|y)] = E_y\left[\frac{p(y|x)p(x)}{p(y)}\right] = p(x)E_y\left[\frac{p(y|x)}{p(y)}\right] = p(x)\sum_y \frac{p(y|x)}{p(y)}p(y) = p(x)\sum_y p(y|x) = p(x)\cdot 1$$

Proof for $\boldsymbol{var[x] = E_y[var_x[x|y]] + var_y[E_x[x|y]]}$:

$$E_y[var_x[x|y]] = E_y[E_x[x^2|y] - E_x[x|y]^2] = E_y\left[\sum_x x^2 p(x|y) - \left(\sum_x xp(x|y)\right)^2\right] =$$

$$\sum_x x^2 \underbrace{E_y[p(x|y)]}_{by\ (*):\ p(x)} - \underbrace{E_y\left[\left(\sum_x xp(x|y)\right)^2\right]}_{A:=} = \sum_x x^2 p(x) - A = \boxed{E[x^2] - A}$$

$$var_y\left[E_x[x|y]\right] = var_y\left[\sum_x xp(x|y)\right] = \underbrace{E_y\left[\left(\sum_x xp(x|y)\right)^2\right]}_{=A} - E_y\left[\sum_x xp(x|y)\right]^2 = A - \left(\sum_x x \underbrace{E_y[p(x|y)]}_{by\ (*):\ p(x)}\right)^2 =$$

$$A - \left(\sum_x xp(x)\right)^2 = \boxed{A - E[x]^2}$$

$$\Rightarrow E_y\left[var_x[x|y]\right] + var_y\left[E_x[x|y]\right] = E[x^2] - A + A - E[x]^2 = E[x^2] - E[x]^2 = \boxed{var[x]}$$

3) Project Proposal: Stylometry using 'Writeprints'

Problem Statement and Project Proposal

Stylometry is the application of the study of linguistic style to authorship attribution. The main domain it is used for is written language – identifying an anonymous author of a text by mining it for linguistic features. Authorship attribution has applications across domains such as literature, history and forensic linguistics, with various examples including authenticating works of Shakespeare, identifying the authors of the famous Federalist Papers and analyzing suicide letters.

One of the novel and most effective methods for authorship attribution is Writeprints [2], presented later in more details. My project proposal is to integrate implementation of this method into JStylo [1], an open source Java-based authorship attribution platform I developed as part of my work at PSAL (The Privacy, Security and Automation lab at Drexel). The platform enables extracting various linguistic features, and currently allows usage of several Weka classifiers.

I intend to evaluate the implementation on the Extended Brennan-Greenstadt Adversarial Corpus [3] which includes documents of 45 different authors with at least 6500 words per author, including adversarial documents where the authors try to change their writing style by 1) attempting to hide it, and 2) imitating another author (Cormac McCarthy).

The evaluation will include both cross-validation on the non-obfuscated documents and evaluating how robust is the Writeprints method in adversarial settings.

Background and Importance

The field of stylometry has naturally become dominated by machine learning approaches, harnessing computational power to the task to achieve high performance and accuracy. Stylometry challenges can be divided into three main sets: identifying an author from a given set of candidates (supervised learning), classifying a set of anonymous documents into clusters where each belongs to a different author (unsupervised learning) and analyzing text for different author features, such as personality traits, age, gender, number of authors of the questioned document etc.

Current research in the field is advancing in three directions:

- Developing stylometry techniques to achieve high accuracy over a varying number of candidate authors, and display better methods of feature selection and cross-domain effectiveness.
- The role of stylometry in privacy and anonymity.
- Adversarial Stylometry – attacking and circumventing stylometry techniques, and countering such attacks.

<u>Stylometry, Privacy and Anonymity</u>

Stylometry has an impact over anonymous communication. As users of the internet may want to hide their identity, other than anonymizing their technical identities (e.g. their source IP address), their writing style also has to be taken under consideration. Keeping the privacy and anonymity of internet users may have great value, especially in cases where anonymous communication is essential to exposing crime, government corruption, human rights abuses etc. Therefore in order to keep a private identity, it is necessary to study stylometric techniques and approaches in order to learn how to circumvent such techniques that may be applied on the published communication. Many may think it is sufficient to cover this aspect only for large text samples, but ongoing research demonstrates stylometry abilities that can be applied for short instant messages, twits and similar.

Approaches

The most common approach to stylometry is supervised analysis of the feature space. Applied in machine learning, this means training a classifier based on a set of known documents of all candidate authors, and applying that classifier to the set of test / unknown documents to determine their author.

When approaching a stylometry task, the main configuration that defines the approach is the set of features examined. The domain of possible features is virtually endless (for instance, choosing English letter n-grams alone generates 26^n features), so it is important to select features that best distinguish between authors carefully. Although it seems difficult, decades of research suggest several features that have been proven effective [4]:

- *Function words*: words used to describe a relationship between meaningful words, and are topic-independent. For instance, articles (“the”, “a”, “an”), pronouns (“he”, “she”, “him”) and particles (“if”, “however”, “thus”).
- *Vocabulary features*: the vocabulary richness of a document, or word selection out of a set of synonyms.
- *Syntactic features*: the grammatical structures used by the author to convey his ideas. For instance, various punctuation features such as frequency of different punctuations across the text.

Methods for classification may include neural networks [5, 6, 7], Support Vector Machines (SVM) [8, 9, 10], Bayesian classifiers, decision trees and others. In addition to the above, there are two unique methods designed specifically for stylometry tasks, and have been shown to be promising:

- *The WritePrints method* [2]: developed by Abassi and Chen, this method incorporates a rich set of features, and uses a sliding window and pattern disruption algorithm with individual author-level feature sets. This method is used for both authorship attribution (supervised) and similarity detection (unsupervised). The method is shown to achieve higher accuracy than known before, across tasks with high number of candidate authors.
- *The Synonym-Based method* [11]: developed by Clark and Hannon, this method measures the selection of each word and weighs that measurement by accounting for how common that word is and how many choices the author had.

References

[1] http://psal.cs.drexel.edu/

[2] A. Abbasi and H. Chen. *Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace*. ACM Trans. Inf. Syst., 26(2):1–29, 2008.

[3] M. Brennan and R. Greenstadt. *Practical Attacks against Authorship Recognition Techniques*. In Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence (IAAI), Pasadena, California, July 2009.

[4] P. Juola. *Authorship attribution*. Foundations and Trends in information Retrieval, 1(3):233–334, 2008.

[5] R. A. J. Matthews and T. V. N. Merriam. *Neural computation in stylometry I: An application to the works of Shakespeare and Marlowe*. Literary and Linguistic Computing, vol. 8, no. 4, pp. 203–209, 1993.

[6] T. V. N. Merriam and R. A. J. Matthews. *Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe*. Literary and Linguistic Computing, vol. 9, no. 1, pp. 1–6, 1994.

[7] F. J. Tweedie, S. Singh, and D. I. Holmes. *Neural network applications in stylometry: The Federalist Papers*. Computers and the Humanities, vol. 30, no. 1, pp. 1–10, 1996.

[8] A. Abbasi and H. Chen. *Applying authorship analysis to extremist-group web forum messages*. IEEE Intelligent Systems, vol. 20, no. 6, pp. 67–75, 2005.

[9] M. Koppel and J. Schler. *Ad-hoc authorship attribution competition approach outline*. In Ad-hoc Authorship Attribution Contest, (P. Juola, ed.), ACH/ALLC 2004, 2004.

[10] R. Zheng, J. Li, H. Chen, and Z. Huang. *A framework for authorship identification of online messages: Writing-style features and classification techniques*. Journal of the American Society for Information Science and Technology, vol. 57, no. 3, pp. 378–393, 2006.

[11] J. H. Clark and C. J. Hannon. *A classifier system for author recognition using synonym-based features*. Lecture Notes in Computer Science, 4827:839–849, 2007.