Exercises

1. In a game of guessing a randomly drawn word from a dictionary (a) what is the amount of information obtained for knowing the answer? (b) compare the amounts of information learning that the answer is "the" and "zoo". (c) compare the amounts of information learning that the answer begins with 't' and 'z'. (d) what is the entropy of the word guess event? (e) if a word is drawn form a corpus rather than a dictionary, compare the amounts of information again for learning that the answer is "the" and "zoo".
2. Multiple choices

(A) self-information

(B) mutual information

(C) PMI

(D) entropy

(E) perplexity

(F) cross-entropy

(G) model perplexity

(H) KL-divergence

(a) which of the above measures concern individual outcomes of random events ?
(b) which of the above measures are event-level measures?

(c) which of the above measures study a single distribution?
(d) which of the above measures study two different distributions?
(e) entropy and _____ offer the same information in different measures.
(f) cross-entropy and _____ offer the same information in different measures.

3. State the correlation between log-linear model and entropy/cross-entropy?

4. Given a training corpus for document classification, can you use PMI between words and class labels to find out which words are the most representative for each class? Compare your results with the weights on feature instances for log-linear classification for the same words.

5. Prove in Eq 5.5 that $\sum_x \sum_y P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} \geq 0$. (Hint: $\log_2 a \leq (a-1)\log_2 e$ inequality state that $\log_2 \frac{P(x,y)}{P(x)P(y)} \geq (1 - \frac{P(x)P(y)}{P(x,y)}) \log_2 e$.)

6. Use t-test as introduced in Chapter 4 to define distributional vector representations of words of a document.

7. Consider again the classification task with unbalanced data. Suppose that we have manually labeled a set of training data $D = \{(x_i, c_i)\}|_{i=1}^N$, the main training objective can be to maximise the likelihood $\sum_{i=1}^N \log P(\mathbf{c}_i|x_i)$. Now suppose that we have balanced samples of each C_i in D, but the distribution of class labels on unseen data is very imbalanced, with $P(\mathbf{c}_i) = \gamma_i$, $i \in [1 \ldots |C|]$. We can integrate this knowledge of class label distribution into our training objective, by measuring the KL divergence of our model-given class distribution and the class distribution $P(\mathbf{c}_i)$. In particular, the model-given class distribution can be estimated by using a current model to label a large set of raw inputs $R = \{x'_i\}|_{i=1}^{N'}$, and then counting the relative frequencies of class labels $Q(\mathbf{c}_i)$ on the outputs. The final training objective, after integrating this knowledge, is $\sum_{i=1}^N \log P(\mathbf{c}_i|x_i) - \lambda D_{KL}\Big(P(\mathbf{c}_i), Q(\mathbf{c}_i)\Big)$, where λ is a hyper parameter to control the importance of the regularisation term.

(a) How can the value of λ be determined?

(b) Can $Q(\mathbf{c}_i)$ be calculated using relative frequencies of $\mathbf{c}_i$ over the model output on R? Why? (Hint: should the regularisation term be a constant or a function of model parameters?)

(c) A reasonable way to calculate $Q(\mathbf{c}_i)$ is to use the *mathematical expectation* $\sum_{x_j \in R} P(\mathbf{c}_i|x'_j) \cdot Q(x'_j, \mathbf{c}_i)$ as $Q(\mathbf{c}_i)$. Now define $Q(x'_j, \mathbf{c}_i)$ in order to use the model score $P(c|x)$ as the only term to denote $Q(\mathbf{c}_i)$. (This method is named **expectation regularisation**.)

(d) Denote a feature vector as $\vec{\phi}(x, c)$, calculate the derivative of the training objective with respect to a model parameter $\vec{\theta}$.

(e) Derive a SGD training algorithm for the training objective above.

8. *Information gain* and χ^2 *statistic* have been used as the criteria for feature selection. Consider a text classifier with bag-of-word features. *Information*

gain of a word w is defined as :

$$\begin{aligned} IG(w) = &- \sum_{i=1}^{n} \hat{P}(c_i) \log \hat{P}(c_i) \\ &+ \Big(\hat{P}(w) \sum_{i=1}^{n} \hat{P}(c_i|w) \log \hat{P}(c_i|w) \\ &+ \hat{P}(\overline{w}) \sum_{i=1}^{n} \hat{P}(c_i|\overline{w}) \log \hat{P}(c_i|\overline{w}) \Big) \end{aligned} \tag{5.20}$$

where c_i are the set of class labels and $\overline{w}$ denotes a word that is not within the training set for (w, c_i). χ^2 statistic is defined as

$$\chi^2(w, c_i) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}, \tag{5.21}$$

where A is the number of times w and c_i co-occur, B is the number of times w occurs with c_i, C is the number of times C_i occurs without w and D is the number of time neither c_i nor w occurs. w can be selected using

$$\chi^2_{avg}(w) = \sum_{i=1}^{n} \hat{P}(c_i) \chi^2(w, c_i) \tag{5.22}$$

or

$$\chi^2_{max}(w) = max^n_{i=1} \chi^2(w, c_i). \tag{5.23}$$

As discussed in Section 5.3.3, using mutual information

$$PMI(w, c_i) = \log_2 \frac{\hat{P}(w, c_i)}{\hat{P}(w)\hat{P}(c_i)} = \frac{A \times N}{(A + B) \times (A + C)}, \tag{5.24}$$

w can be selected using

$$PMI_{avg}(w) = \sum_{i=1}^{n} \hat{P}(c_i) \times PMI(w, c_i) \tag{5.25}$$

or

$$PMI_{max}(w) = max^n_{i=1} PMI(w, c_i) \tag{5.26}$$

Discuss the correlation between $TF(w) \cdot IDF(w)$, $IG(w)$, $\chi^2(w)$ and $PMI(w)$ for feature selection. What are the similarities? What can be their relative strength? Compare them empirically for SVM classification.