Spamerkennung mit log-linearen Modellen

Implementieren Sie ein weiteres Spamerkennungs-System auf Basis von log-linearen Modellen, welches dieselben Daten wie in der letzten Übung verwendet. Es sollte wieder aus einem Trainingsprogramm und einem Anwendungsprogramm bestehen. Ihr Programm sollte so allgemein implementiert werden, dass es auch für Klassifikations-Probleme mit mehr/anderen Klassen anwendbar ist.

Hier ist Pseudocode für die Aufgabe:

für n Epochen
    Trainingsdaten zufällig umordnen mit random.shuffle
    für jede Mail in den Trainingsdaten
        Merkmalsvektoren für alle Klassen berechnen
        Scores für alle Klassen berechnen
        p(Klasse|Mail) für alle Klassen berechnen
        Gradient durch Addition der beobachteten Merkmalswerte und
            Subtraktion der erwarteten Merkmalswerte berechnen
        Gewichtsvektor anpassen

Die Programme werden folgendermaßen aufgerufen:

```
python3 train.py train-dir paramfile
```

```
python3 classify.py paramfile mail-dir
```

Zur Vermeidung von **Overflow** gehen Sie so vor:

- Sie berechnen erst für jede Klasse c den Score $s(c)$ als Produkt von Merkmals- und Gewichtsvektor.
- Dann berechnen Sie $logZ = \log \sum_c e^{s(c)} = \text{logsumexp}_c\ s(c)$
- Zuletzt berechnen Sie $p(c) = \frac{e^{s(c)}}{Z} = \frac{e^{s(c)}}{e^{\log Z}} = e^{s(c) - logZ}$

Sie müssen keine Regularisierung machen.

Log-Sum-Exp-Trick

Für $m := \max_x x$ gilt:

$$\log \sum_x e^x = \left(\log \sum_x e^x\right) - \log e^m + m = \left(\log \sum_x \frac{e^x}{e^m}\right) + m = \left(\log \sum_x e^{x-m}\right) + m$$

Die Werte e^{x-m} liegen zwischen 0 und 1 und können (anders als e^x) keinen Overflow verursachen.

Vorüberlegungen

- Welche Merkmale verwenden Sie am besten?

- Welche Datenstrukturen verwenden Sie?
- Was speichern Sie in der Parameterdatei?
- Welche Teilaufgaben umfasst das Anwendungsprogramm?

Wenn Sie wollen, können Sie die Development-Daten verwenden, um Hyperparameter wie die Lernrate zu optimieren.

Schicken Sie das fertige **Programm** und die Liste der für die Testdaten **ausgegebenen Klassen** an `schmid@cis.lmu.de`.