9. **Смесено обучение(с и без учител)**

• **Semisupervised обучение**
• Клъстериране за класификация
• **Cotraining**
• **EM** и **cotraining**
• Подходи с невронни мрежи
• **Многоинстанционно обучение**
• Преобразуване в едноинстанционно обучение
• Подобряване на алгоритмите за обучение
• Специализирани методи за многоинстанционно обучение

Semisupervised обучение

• **Semisupervised обучение** се опитва да използва както немаркирани, така и маркирани данни.
• Целта е да подобри производителността на класификацията.
• Защо да правим това? Защото немаркираните данни често са в изобилие, а маркирането на данни може да бъде скъпо.
• **Web mining**: класифициране на уеб страници.
• **Text mining**: идентифициране на имена в текст.
• **Video mining**: класифициране на хора в новините.
• Използването на големия набор от немаркирани примери би било много привлекателно.

Клъстериране за класификация

• Видяхме как да използваме **EM** за обучение на смесен модел за клъстериране, с една смесена компонента на клъстер.
• **Naïve Bayes** може да се разглежда като прилагане на смесен модел с една компонента на разпределение за клас.
• Можем ли да комбинираме двете?
• Идея: използване на **Naïve Bayes** върху маркирани примери и след това прилагане на **EM**.

• Първо, изграждане на модел **Naïve Bayes** върху маркираните данни.
• Второ, маркиране на немаркираните данни въз основа на вероятностите за класовете (**expectation** стъпка).
• Трето, обучение на нов модел **Naïve Bayes** въз основа на всички данни (**maximization** стъпка).
• Четвърто, повтаряне на втората и третата стъпка до сходимост.
• По същество е същото като **EM** за клъстериране с фиксирани вероятности за членство в клъстери за маркираните данни.

Коментари по този подход
• Предполага условна независимост.
• Успешно е приложен при класификация на документи:
• Някои фрази са индикативни за класовете.
• Някои от тези фрази се срещат само в немаркираните данни, други — и в маркираните, и в немаркираните данни.
• **EM** може да обобщи модела отвъд маркираните данни, като се възползва от съвместните срещания на тези фрази.
• Усъвършенстване 1: намаляване на теглото на немаркираните данни.
• Въвеждане на параметър, който позволява на потребителя да дава по-малко тегло на немаркираните данни по време на процеса на обучение.
• Усъвършенстване 2: позволяване на множество клъстери за клас.
• Можем да разширим смесения модел, за да има множество компоненти за клас, а не само една.
• Модифициране на стъпката за максимизация, за да не само вероятностно маркира всеки пример с класове, но и да го присвоява вероятностно към компоненти в рамките на клас.

Cotraining

• **Cotraining** е друг добре известен метод за s**emisupervised обучение**.

• Използва множество изгледи (**multiple views**, множество набори от атрибути) за обучение от маркирани и немаркирани данни.

• Уеб страниците са класически пример за данни с множество изгледи:

• Първият набор от атрибути описва съдържанието на уеб страницата.

• Вторият набор от атрибути описва връзките, които сочат към уеб страницата.

• Алгоритъм за **cotraining**:

• Стъпка 1: изграждане на класификационен модел от всеки изглед.

• Стъпка 2: използване на моделите за присвояване на етикети на немаркираните данни.

• Стъпка 3: избор на тези немаркирани примери, които бяха предсказани с най-голяма увереност (идеално, запазвайки съотношението на класовете).

• Стъпка 4: добавяне на тези примери към тренировъчния набор.

• Стъпка 5: връщане към стъпка 1, докато немаркираните данни не бъдат изчерпани.

• Предположение: изгледите са независими (но **cotraining** изглежда работи и когато изгледите са зависими).

Комбиниране на EM и cotraining

• Можем да комбинираме **EM** и **cotraining** за **Semisupervised обучение**, за да получим метода **co-EM**.

• Работи като основния подход на **EM**, но изгледът/класификаторът се сменя при всяка итерация на **EM**.

• Използва всички немаркирани екземпляри, претеглени чрез оценките на вероятностите за класовете на класификаторите, за

обучение при всяка итерация.
• Основният метод **cotraining** присвоява твърди етикети вместо това.
• Методът **co-EM** е използван успешно и с машини с поддържащи вектори (**support vector machines**), адаптирани да работят с тегла.
• Логистични модели се напасват към изхода на SVM, за да се получат оценки на вероятностите за класовете.
• **Cotraining** и **co-EM** изглежда работят дори когато изгледите са избрани случайно.
• Защо? Вероятно защото котренираният класификатор е по-робустен.

Подходи с невронни мрежи
• **Semisupervised обучение** може да се приложи и при **дълбоко обучение** на класификатори с невронни мрежи.
• Ненадзираното предварително обучение (**unsupervised pre-training**) е форма на **Semisupervised обучение** в **дълбокото обучение**.
• Чисто supervised **дълбоко обучение** е много ефективно, когато са налични големи количества маркирани данни.
• **Unsupervised pre-training** въз основа на немаркирани данни може да бъде полезно, когато маркираните данни са малко.
• Също така е възможно да се разширят автокодерите (**autoencoders**), които са ненадзирани, за да включат надзор, когато е наличен.
• Добавяне на клон към изходния слой на автокодера, който предсказва етикета на класа.
• Прилагане на комбинирана функция за загуба (**composite loss function**), която измерва както производителността при възстановяване, така и производителността при класификация.