Машинное обучение (Machine Learning)

7 приемов для более эффективного обучения НС

Уткин Л.В.

Санкт-Петербургский политехнический университет Петра Великого

Что такое эффективное обучение (1)

- Создание и обучение НС требует множества параметров (количество и типы узлов, количество слоев, скорость обучения, обучающая и тестовая выборки и т.д.).
- Нет четкого правила для выбора, так как они зависят от данных.
- Две проблемы одновременно:
  1. **Обучение** на обучающих данных, чтобы минимизировать функцию потерь.
  2. **Обобщение** - прогнозирование на новых примерах.

Что такое эффективное обучение (2)

- Две проблемы одновременно:
  1. **Обучение** на обучающих данных, чтобы минимизировать функцию потерь.
  2. **Обобщение** - прогнозирование на новых примерах.
- Компромисс между проблемами: модель, которая учится слишком хорошо, будет плохо обобщать, а модель, которая хорошо обобщает, может оказаться недостаточно обученной.

Смещение и дисперсия

Проблема обучения НС рассматривается с точки зрения компромисса между смещением (bias) и дисперсией (variance):

- **Смещение**: мера того, как выход НС, усредненный по всему обучающему множеству, отличается от того, что мы хотим
- **Дисперсия**: мера того, насколько выход НС варьируется для разных данных

Как долго учить?

От модели с большим смещением и малой дисперсией в начале обучения к модели с более низким смещением и более высокой дисперсией в конце обучения.

Если обучать НС слишком долго, то она начинает учиться на шуме, который имеется в данных - переобучение. Дисперсия будет большой, потому что шум варьируется между данных.

7 приемов

1. Stochastic Versus Batch Learning
2. Shuffling the Examples
3. Normalizing the Inputs
4. The Sigmoid
5. Choosing Target Values
6. Initializing the Weights
7. Choosing Learning Rates

Stochastic Versus Batch Learning

Компромис между стохастическим и пакетным градиентным спуском:

- **Пакетный (batch)**, когда на каждой итерации обучающая выборка просматривается целиком и затем веса модифицируются.
- **Стохастический (stochastic, online)**, когда на каждой итерации из обучающей выборки каким-то (случайным) образом выбирается только один объект.

Stochastic Learning и Mini-Batch

- Обычно намного быстрее, чем пакетное обучение
- Часто приводит к лучшим решениям
- Может быть использовано для отслеживания изменений.

Но шум может все испортить! Чтобы уменьшить флуктуации, можно либо уменьшить скорость обучения, либо сделать адаптивным размер пакета.
Выход: **Mini-Batch Gradient Descent**

Shuffling the Examples (Перемешивание)

НС обучается быстрее на наиболее неожиданных примерах

- Предлагается случайно выбирать объекты, но попеременно из разных классов. Идея в том, что объекты из разных классов скорее всего менее "похожи чем объекты из одного класса, поэтому вектор весов будет каждый раз сильнее изменяться.
- Возможен вариант алгоритма, когда выбор каждого объекта неравновероятен, причҷм вероятность выпадения объекта обратно пропорциональна величине ошибки на объекте.

Normalizing the Inputs (Нормализация)

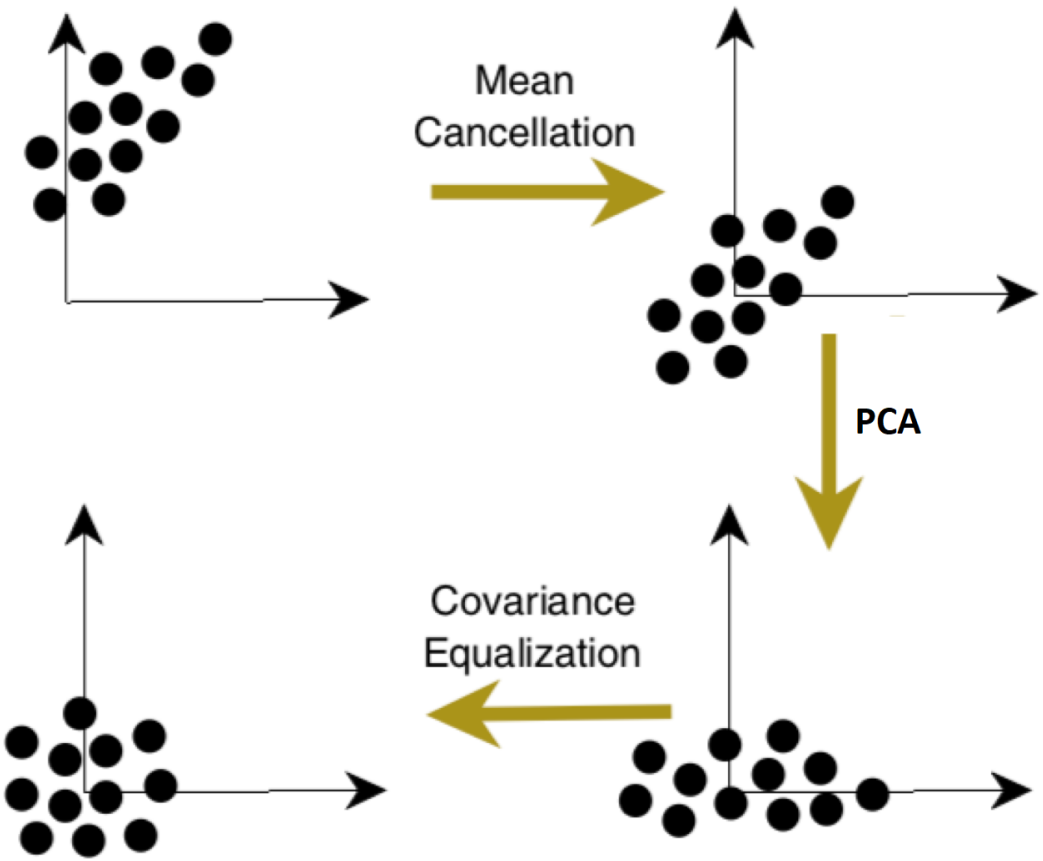
Сходимость обычно быстрее, если среднее значение каждой входной переменной из обучающего множества близко к нулю

Эту эвристику следует применять на всех слоях НС, т.е. среднее значение выходных данных узла должно быть близко к нулю, т.к. эти выходные данные являются входными данными для следующего слоя

Масштабирование ускоряет обучение, т.к. помогает сбалансировать скорость, с которой модифицируются веса - СКО=1

Декорреляция входных данных - устранение любой линейной зависимости между входными переменными (методом главных компонент)

Нормализация



Sigmoid

Симметричные сигмоиды (гиперболический тангенс) часто сходятся быстрее, чем стандартная логистическая функция.

Choosing Target Values

- Что выбрать {-1,1} или {0,1}?
- Здравый смысл подсказывает, что целевые значения должны быть установлены на значение асимптот сигмоиды. **Но**...
  - Для достижения значений в крайних точках сигмоиды могут потребоваться большие веса, что сделает модель нестабильной
- Целевые значения - в точке максимальной второй производной на сигмоиде.
  - {0.1,0.9} вместо {0,1}

Initializing the Weights

Веса должны выбираться случайным образом, но таким образом, чтобы сигмоид сначала активировался в своей линейной области. Тоже самое и для ReLU, где линейная часть функции положительна.

Почему:

1. градиенты достаточно велики, чтобы обучение могло продолжаться
2. сеть изучит линейную часть отображения перед более сложной нелинейной частью

Choosing Learning Rates

- Уменьшают скорость обучения, когда весовой вектор «колеблется», и увеличивают, когда он устойчив
- Различная скорость обучения для каждого веса может улучшить сходимость
- Скорость обучения должна быть пропорциональна квадратному корню из числа входов в нейрон
- Веса в нижних слоях обычно должны быть больше, чем в более высоких слоях

Вопросы

?