# О задаче обучения смеси экспертов

## 1 Пока это поток мыслей о генерации данных

Основной задачей машинного обучения является *восстановление зависимостей* по некоторым *данным*.

Для восстановления зависимостей нужны данные. Также важно понимать, что это за данные, откуда они, как они порождены.

Для восстановления зависимостей обычно вводятся гипотезы о том как данные были порождены. Рассмотрим следующие гипотезы порождения данных:

1. **Базовая гипотеза порождения данных** — гипотеза порождения данных, которая состоит из признакового описания объектов, отображения объектов в некоторые ответы, ответы для заданных объектов.

2. **Гипотеза порождения данных со смесью моделей** — гипотеза порождения данных, в которой вместо одной модели используется несколько моделей, причем выбор модели для каждого объекта происходит не зависимо от самого объекта.

3. **Гипотеза порождения данных со смесью экспертов** — гипотеза порождения данных, в которой вместо одной модели используется несколько моделей, причем модель для каждого объекта выбирается на основе самого объекта.

Далее гипотезу порождения данных разберем на основе предположений о генерации этих данных природой. А нашей итоговой задачей будет понять, как природа сгенерировала данные.

### 1.1 Базовая гипотеза порождения данных

Разберем базовую гипотезу порождения данных более подробно. Для простоты рассмотрим порождения данных, где в качестве целевой переменной будет число 0 или 1 — задача классификации.

Процесс порождения данных природой следующий:

1. Пусть природе изначально выдали объекты, причем задали признаковое описание объектов:

$$\mathbf{X} \in \mathbb{R}^{m \times n},$$

где $m$ — число объектов, а $n$ — число признаков.

2. Далее природа случайным образом выбирает некоторый вектор:

$$\mathbf{w} \sim p\big(\mathbf{w}\big)$$

где $p\big(\mathbf{w}\big)$ плотность некоторого распределения, $\mathbf{w} \in \mathbb{R}^n$.

3. Далее природа для каждого объекта $\mathbf{x}_i$ вычисляет его класс следующим образом:

$$y_i \sim \mathrm{Be}\big(\sigma\big(\mathbf{x}_i^\mathsf{T}\mathbf{w}\big)\big),$$

где $\sigma\big(t\big) = \frac{1}{1+e^{-t}}$.

К примеру мы вводим такую гипотезу порождения данных, когда начинаем решать задачу при помощи логистической регрессии. Разные методы машинного обучения рассматривают свои гипотезы порождения данных на 3-м этапе. Но 1-й и 2-й пункт обычно не меняется.

## 1.2 Гипотеза порождения данных смеси моделей

Усложним генерацию данных тем, что теперь у нас модель не единственная, то есть скорректируем второй шаг. Также рассмотрим задачу классификации на два класса.

Процесс порождения данных природой следующий:

1. Пусть природе изначально выдали объекты, причем задали признаковое описание объектов:

$$\mathbf{X} \in \mathbb{R}^{m\times n},$$

где $m$ — число объектов, а $n$ — число признаков.

2. Далее природа случайным образом выбирает не один, а несколько векторов (к примеру два вектора):

$$\mathbf{w}_1 \sim p_1\big(\mathbf{w}\big), \quad \mathbf{w}_2 \sim p_2\big(\mathbf{w}\big),$$

где $p_1\big(\mathbf{w}\big)$ и $p_1\big(\mathbf{w}\big)$ некоторые плотности распределения, причем $\mathbf{w} \in \mathbb{R}^n$.

3. Далее природа для каждого объекта $\mathbf{x}_i$ вычисляет его класс следующим образом:

$$k \sim \mathrm{Cat}\big([0.5, 0.5]\big), \quad y_i \sim \mathrm{Be}\big(\sigma\big(\mathbf{x}_i^\mathsf{T}\mathbf{w}_k\big)\big),$$

где $\sigma\big(t\big) = \frac{1}{1+e^{-t}}$.

То есть основное отличие от первой гипотезы в том, что в гипотезе порождения данных фигурирует не одна модель, а несколько. Но мы также вводим предположения, что природа выбирая модель не смотрела на сам объект.

### 1.3 Гипотеза порождения данных смеси экспертов

Далее рассмотрим вариант генерации данных, когда природа при выборе модели будет учитывать сам объект для которого выбирается модель.

Процесс порождения данных природой следующий:

1. Пусть природе изначально выдали объекты, причем задали признаковое описание объектов:
$$\mathbf{X} \in \mathbb{R}^{m \times n},$$
где $m$ — число объектов, а $n$ — число признаков.

2. Далее природа случайным образом выбирает не один, а несколько векторов (к примеру два вектора):
$$\mathbf{w}_1 \sim p_1\big(\mathbf{w}\big), \quad \mathbf{w}_2 \sim p_2\big(\mathbf{w}\big),$$
где $p_1\big(\mathbf{w}\big)$ и $p_1\big(\mathbf{w}\big)$ некоторые плотности распределения, причем $\mathbf{w} \in \mathbb{R}^n$.

3. Далее природа для каждого объекта $\mathbf{x}_i$ вычисляет его класс следующим образом:
$$k = \pi\big(\mathbf{x}_i\big), \quad y_i \sim \mathrm{Be}\big(\sigma\big(\mathbf{x}_i^\mathsf{T}\mathbf{w}_k\big)\big),$$
где $\sigma\big(t\big) = \frac{1}{1+e^{-t}}$, а $\pi$ — некоторое решающее правило природы, по которой для каждого объекта производится выбор модели.

Заметим, что отличие смеси экспертов от смеси моделей в том, что для смеси экспертов для каждого объекта модель выбирается не случайным образом, а при помощи некоторого правила.

## 2 Еще не много про генерацию данных

Заметим, что в случае гипотез порождения данных для смесей получаем, что признаковое описания у разных моделей может быть разным.

Введем обобщения, которое состоит в том, что у каждой модели свое признаковое описание объекта.

Про это можно прочитать в другом файле Grabovoy2019ExpertLearningTask.pdf.

Также стоит заметить, что во всех случаях каждая из моделей не обязательно должна быть линейной. В общем случае модели могут иметь разную сложность, к примеру одна из моделей может быть линейной, а другая нейросеть.

## 3 Пример задач о смеси экспертов

Рассмотрим задачу о переводе текста.

Пусть объекты это предложения на разных языках (к примеру на русском и на английском), и все предложения находятся в одной выборке. Для перевода данных предложений требуется решить две подзадачи: определения языка, перевод. Да, данные задачи можно решать по отдельности, но можно решать задачи в единой процедуре.