Оценка параметров многомерного нормального распределения с помощью нейронных сетей

Байесовский вывод, Теория Лоссов и Экспериментальное подтверждение

Исследование свойств NN

28 ноября 2025 г.

1. Постановка задачи

Дано: Мы имеем дело с совместным нормальным распределением вектора признаков X и вектором параметров θ:

$$Z = \begin{pmatrix} X \\ \theta \end{pmatrix} \sim \mathcal{N}(\mu_{joint}, \Sigma_{joint})$$

Параметры заданы явно через блочные матрицы:

$$\mu_{joint} = \begin{pmatrix} \mu_x \\ \mu_\theta \end{pmatrix}, \quad \Sigma_{joint} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{x\theta} \\ \Sigma_{\theta x} & \Sigma_{\theta\theta} \end{pmatrix}$$

Цель: Обучить нейронные сети аппроксимировать параметры **условного распределения** $P(\theta|X)$, не используя аналитические формулы при обучении:

- $\mathbb{E}[\theta|X]$ (Условное среднее)
- $\mathrm{Var}[\theta|X]$ (Условная дисперсия)
- $Q_\rho[\theta|X]$ (Условный квантиль)

2. Почему Байесовский подход?

Нам нужно найти $P(\theta|X)$.
Согласно Теореме Байеса:

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)}$$

Ключевая идея

Знаменатель $P(X)$ не зависит от θ. Следовательно:

$$P(\theta|X) \propto P(X,\theta)$$

Стратегия: Мы возьмем плотность **совместного распределения** (которую мы знаем) и рассмотрим её как функцию от θ при фиксированном X. Это позволит нам найти параметры условного распределения методом выделения полного квадрата.

3. Совместная плотность (Показатель экспоненты)

Плотность совместного нормального распределения имеет вид:

$$P(X, \theta) \propto \exp\left(-\frac{1}{2}(Z - \mu_{joint})^T \Sigma_{joint}^{-1}(Z - \mu_{joint})\right)$$

Обозначим матрицу $\Lambda = \Sigma_{joint}^{-1}$ и разобьем её на блоки:

$$\Lambda = \begin{pmatrix} \Lambda_{xx} & \Lambda_{x\theta} \\ \Lambda_{\theta x} & \Lambda_{\theta\theta} \end{pmatrix}$$

Тогда показатель экспоненты $J(\theta)$ (отбрасывая члены без θ) равен:

$$J(\theta) = (\theta - \mu_\theta)^T \Lambda_{\theta\theta}(\theta - \mu_\theta) + 2(\theta - \mu_\theta)^T \Lambda_{\theta x}(X - \mu_x)$$

4. Выделение полного квадрата (Шаг 1: Дисперсия)

Мы хотим привести $J(\theta)$ к каноническому виду условной гауссианы:

$$(\theta - \mu_{\theta|x})^T \mathbf{\Sigma}_{\theta|\mathbf{x}}^{-1} (\theta - \mu_{\theta|x})$$

Сравним квадратичные члены по θ:

- В нашей формуле: $\theta^T \mathbf{\Lambda}_{\theta\theta} \theta$
- В канонической: $\theta^T \mathbf{\Sigma}_{\theta|\mathbf{x}}^{-1} \theta$

Результат для Дисперсии

$$\Sigma_{\theta|x} = (\Lambda_{\theta\theta})^{-1} = \Sigma_{\theta\theta} - \Sigma_{\theta x} \Sigma_{xx}^{-1} \Sigma_{x\theta}$$

Вывод: В формуле **НЕТ** X. Условная дисперсия — константа.

5. Выделение полного квадрата (Шаг 2: Среднее)

Теперь сравним линейные члены по θ:

- В нашей формуле: $2\theta^T \Lambda_{\theta x}(X - \mu_x)$
- В канонической: $-2\theta^T \Sigma_{\theta|x}^{-1} \mu_{\theta|x}$

Приравниваем и выражаем $\mu_{\theta|x}$:

$$\Sigma_{\theta|x}^{-1} \mu_{\theta|x} = -\Lambda_{\theta x}(X - \mu_x) + \dots$$

После алгебраических преобразований:

Результат для Среднего

$$\mu_{\theta|x} = \mu_\theta + \Sigma_{\theta x} \Sigma_{xx}^{-1}(X - \mu_x)$$

Вывод: Зависимость линейная ($Ax + b$).

6. Условный Квантиль

Так как условное распределение $\theta|X$ является нормальным, его квантиль Q_ρ однозначно определяется через среднее и дисперсию.

$$Q_\rho[\theta|X] = \mu_{\theta|x} + \text{std}_{\theta|x} \cdot \Phi^{-1}(\rho)$$

Подставляя наши результаты:

$$Q_\rho[\theta|X] = \underbrace{[Ax + b]}_{\text{Mean}} + \underbrace{\sqrt{Const} \cdot z_\rho}_{\text{Сдвиг}}$$

Вывод: Условный квантиль — это тоже линейная функция от X, параллельная функции среднего.

7. Обоснование Loss-функций: Среднее (L2)

Почему мы используем MSE для предсказания среднего?
Рассмотрим задачу минимизации с L2 лоссом:

$$\hat{c} = \arg\min_{c} \mathbb{E}_Y[(Y - c)^2]$$

Возьмем производную по c и приравняем к нулю:

$$\frac{\partial}{\partial c}\mathbb{E}[Y^2 - 2Yc + c^2] = \mathbb{E}[-2Y + 2c] = 0$$

$$-2\mathbb{E}[Y] + 2c = 0 \implies c = \mathbb{E}[Y]$$

Итог: Нейросеть с L2 лоссом математически обязана сходиться к условному матожиданию.

8. Обоснование Loss-функций: Дисперсия

Как научить сеть предсказывать дисперсию? Мы знаем, что $\mathrm{Var}(Y) = \mathbb{E}[(Y-\mu)^2]$. Если μ известно (или мы используем его оценку), мы можем определить новую целевую переменную $Z = (Y-\mu)^2$. Тогда задача сводится к предсказанию среднего для Z:

$$\min_v \mathbb{E}[(Z-v)^2] \implies v = \mathbb{E}[Z] = \mathbb{E}[(Y-\mu)^2] = \mathrm{Var}(Y)$$

9. Обоснование Loss-функций: Квантиль (Pinball)

Для квантиля уровня ρ используется Pinball Loss:

$$\mathcal{L}_\rho(u) = \begin{cases} \rho u & u \geq 0 \\ (\rho - 1)u & u < 0 \end{cases}, \quad u = y - \hat{y}$$

Возьмем производную по предсказанию $\hat{y}$ (в точке оптимума матожидание градиента = 0):

$$\mathbb{E}[\nabla \mathcal{L}] = -\rho \cdot P(y > \hat{y}) + (1 - \rho) \cdot P(y \leq \hat{y}) = 0$$

$$\rho(1 - P(y \leq \hat{y})) = (1 - \rho)P(y \leq \hat{y})$$

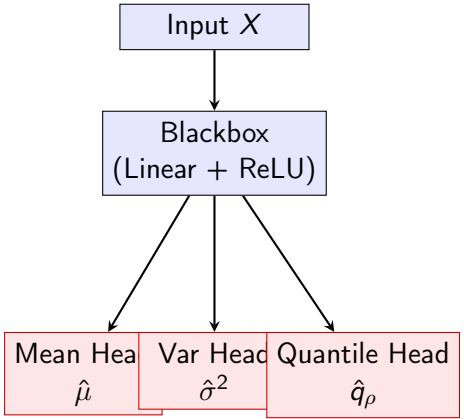
$$P(y \leq \hat{y}) = \rho$$

Итог: Минимум достигается ровно тогда, когда предсказание $\hat{y}$ является ρ-квантилем.

10. Архитектура Нейронной Сети

Компоненты:

- **Blackbox:** Общий энкодер (MLP, Hidden=16).
- **Mean Head:** Линейный слой (MSE Loss).
- **Var Head:** Линейный слой + **Softplus** (для $v > 0$).
- **Quantile Head:** Линейный слой (Pinball Loss).



11. Дизайн Эксперимента

Мы проверяем гипотезу о независимости дисперсии, анализируя **дисперсию предсказаний** $\mathrm{Var}(f(X_{test}))$ при увеличении размера обучающей выборки N.

Цель	**Истинная зависимость**	**Ожидаемое поведение**
Mean	Линейная $(Ax + b)$	$\mathrm{Var}(pred) \uparrow$ (Рост)
Quantile	Линейная $(Ax + b)$	$\mathrm{Var}(pred) \uparrow$ (Рост)
Variance	**Константа**	$\mathrm{Var}(pred) \downarrow$ (Падение к 0)

12. Результаты: Динамика обучения

Наблюдения из графиков:

- **Mean, Quantile:** На малых выборках модели предсказывают константу (недообучение). С ростом N они "хватают"линейный тренд $\Longrightarrow$ дисперсия предсказаний растет.
- **Variance:** На малых выборках модель переобучается под шум (высокая дисперсия). С ростом N она сходится к истинной константе $\Longrightarrow$ дисперсия предсказаний падает.

Это первое экспериментальное доказательство того, что $\Sigma_{\theta|X}$ не зависит от X.

13. Доказательство 2: Анализ весов (L1)

Мы применили L1-регуляризацию (Lasso), чтобы принудительно обнулить незначимые веса матрицы A (первого слоя).

Результаты (L1-норма):

Модель	**Норма ($N = 32$)**	**Норма ($N = 160k$)**
Mean Model	0.29	**0.11** (Высокая)
Variance Model	0.29	**0.0007** (≈ 0)

Интерпретация: В модели среднего веса сохранились (они нужны). В модели дисперсии веса "исчезли". Сеть физически отключила входы X.