

Байесовы нейронные сети

Басов Дмитрий Константинович

1 Обозначения и сокращения

$N(\mu, \sigma^2)$ — нормальное распределение

$\mathbf{x} \cdot \mathbf{y}$ — поэлементное произведение (произведение Адамара)

$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_j (x_j \cdot y_j)$ — скалярное произведение

\mathcal{L} — Evidence Lower Bound (ELBO)

$KL(q||p) = \int q(\mathbf{Z}) \cdot \ln \frac{q(\mathbf{Z})}{p(\mathbf{Z})} d\mathbf{Z}$ — расстояние Кульбака — Лейблера

\mathbf{x} — вектор признаков

\mathbf{y} — таргет

D — датасет — пары значений $\{\mathbf{x}_i, \mathbf{y}_i\}$, где $i = 1, \dots, L$

\mathbf{W} — параметры модели — случайная величина размерности M

$p(D|\mathbf{W}) = \prod_{i=1}^L p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{W})$ — правдоподобие (likelihood)

$p(\mathbf{W})$ — априорное распределение параметров модели (prior)

$p(\mathbf{W}|D)$ — апостериорное распределение параметров модели (posterior)

$p(D)$ — маргинальная вероятность датасета (evidence)

$q(\mathbf{W})$ — аппроксимация апостериорного распределения параметров модели

$p(\mathbf{W}, D) = p(D|\mathbf{W}) \cdot p(\mathbf{W}) = p(\mathbf{W}|D) \cdot p(D)$ — совместная вероятность параметров и данных

2 Постановка задачи

Постановка задачи следующая: у нас есть датасет D и наша цель — смоделировать распределение $p(\mathbf{y}|\mathbf{x}, D)$. То есть мы хотим получить распределение вероятностей таргета \mathbf{y} для неразмеченных \mathbf{x} , используя датасет D . Сделаем следующие преобразования:

$$p(\mathbf{y}|\mathbf{x}, D) = \int p(\mathbf{y}, \mathbf{W}|\mathbf{x}, D) d\mathbf{W} = \int p(\mathbf{y}|\mathbf{W}, \mathbf{x}, D) \cdot p(\mathbf{W}|\mathbf{x}, D) d\mathbf{W} = \int p(\mathbf{y}|\mathbf{W}, \mathbf{x}) \cdot p(\mathbf{W}|D) d\mathbf{W}$$

Получим выражение для $p(\mathbf{W}|D)$, используя формулу Байеса:

$$p(\mathbf{W}|D) = \frac{p(\mathbf{W}, D)}{p(D)} = \frac{p(\mathbf{W}, D)}{\int p(\mathbf{W}, D) d\mathbf{W}} = \frac{p(D|\mathbf{W}) \cdot p(\mathbf{W})}{\int p(D|\mathbf{W}) \cdot p(\mathbf{W}) d\mathbf{W}}$$

Для аппроксимации распределения ответов модели можно воспользоваться методом Монте — Карло: взять сэмплы весов \mathbf{W} из $p(\mathbf{W}|D)$, прогнать их через модель и получить $\hat{\mathbf{y}}$. Однако для этого необходимо уметь сэмплировать из распределения $p(\mathbf{W}|D)$.

Получить аналитическое решение можно только в очень ограниченном числе случаев. Существует возможность сэмплировать из $p(\mathbf{W}|D)$, используя методы Монте — Карло для марковских цепей (MCMC). Однако для больших датасетов и большого числа параметров это становится технически сложно. Альтернативный подход к решению таких задач — аппроксимация распределения $p(\mathbf{W}|D)$ распределением $q(\mathbf{W})$, из которого сэмплировать намного проще.

3 Вариационный вывод для нейронной сети

Запишем выражение ELBO для распределения $q(\mathbf{W})$ и преобразуем его, используя тождество $p(\mathbf{W}, D) = p(\mathbf{W}|D) \cdot p(D)$:

$$\begin{aligned}\mathcal{L}(q(\mathbf{W})) &= \int q(\mathbf{W}) \cdot \ln \frac{p(\mathbf{D}, \mathbf{W})}{q(\mathbf{W})} d\mathbf{W} = \int q(\mathbf{W}) \cdot \ln \frac{p(\mathbf{W}|D) \cdot p(D)}{q(\mathbf{W})} d\mathbf{W} = \ln p(D) \cdot \int q(\mathbf{W}) d\mathbf{W} - \\ &\int q(\mathbf{W}) \cdot \ln \frac{q(\mathbf{W})}{p(\mathbf{W}|D)} d\mathbf{W} = \ln p(D) - KL(q(\mathbf{W})||p(\mathbf{W}|D))\end{aligned}$$

Из равенства $\mathcal{L}(q(\mathbf{W})) = \ln(p(D)) - KL(q(\mathbf{W})||p(\mathbf{W}|D))$ видно, что максимизируя $\mathcal{L}(q(\mathbf{W}))$, мы не только максимизируем $\ln p(D)$, но и минимизируем $KL(q(\mathbf{W})||p(\mathbf{W}|D))$. То есть распределение $q(\mathbf{W})$ будет приближаться к распределению $p(\mathbf{W}|D)$.

Будем максимизировать $\mathcal{L}(q(\mathbf{W}))$. Преобразуем выражение для $\mathcal{L}(q(\mathbf{W}))$, используя тождество $p(\mathbf{W}, D) = p(D|\mathbf{W}) \cdot p(\mathbf{W})$:

$$\begin{aligned}\mathcal{L}(q(\mathbf{W})) &= \int q(\mathbf{W}) \cdot \ln \frac{p(\mathbf{D}, \mathbf{W})}{q(\mathbf{W})} d\mathbf{W} = \int q(\mathbf{W}) \cdot \ln \frac{p(D|\mathbf{W}) \cdot p(\mathbf{W})}{q(\mathbf{W})} d\mathbf{W} = \int q(\mathbf{W}) \cdot \ln p(D|\mathbf{W}) d\mathbf{W} - \\ &\int q(\mathbf{W}) \cdot \ln \frac{q(\mathbf{W})}{p(\mathbf{W})} d\mathbf{W} = \int q(\mathbf{W}) \cdot \ln p(D|\mathbf{W}) d\mathbf{W} - KL(q(\mathbf{W})||p(\mathbf{W}))\end{aligned}$$

Для дальнейшего вывода положим, что распределения $p(\mathbf{W})$ и $q(\mathbf{W})$ являются нормальными с диагональными матрицами ковариации:

$$p(\mathbf{W}) = N(\mathbf{W}|\mathbf{0}, \sigma_{p(\mathbf{W})}^2 \cdot \mathbf{I}), \text{ где } \sigma_{p(\mathbf{W})} \text{ — вектор длины } M$$

$$q(\mathbf{W}) = N(\mathbf{W}|\boldsymbol{\mu}, \sigma_{q(\mathbf{W})}^2 \cdot \mathbf{I}), \text{ где } \boldsymbol{\mu} \text{ и } \sigma_{q(\mathbf{W})} \text{ — вектора длины } M$$

Так как распределения $p(\mathbf{W})$ и $q(\mathbf{W})$ являются нормальными, то $KL(q(\mathbf{W})||p(\mathbf{W}))$ можно посчитать аналитически:

$$KL(q(\mathbf{W})||p(\mathbf{W})) = \frac{1}{2} \sum_{k=1}^M \left(\frac{\sigma_{q(W)_k}^2}{\sigma_{p(W)_k}^2} + \frac{\mu_k^2}{\sigma_{p(W)_k}^2} - \ln \frac{\sigma_{q(W)_k}^2}{\sigma_{p(W)_k}^2} - 1 \right)$$

Априорное распределение параметров модели определяется параметром $\sigma_{p(\mathbf{W})}$. Воспользуемся техникой эмпирического Байеса — нахождения параметров априорного распределения из данных. Посчитаем $\frac{d\mathcal{L}(q(\mathbf{W}))}{d(\sigma_{p(W)_k}^{-2})}$:

$$\begin{aligned}\frac{d\mathcal{L}(q(\mathbf{W}))}{d(\sigma_{p(W)_k}^{-2})} &= \frac{d(\int q(\mathbf{W}) \cdot \ln p(D|\mathbf{W}) d\mathbf{W} - KL(q(\mathbf{W})||p(\mathbf{W})))}{d(\sigma_{p(W)_k}^{-2})} = -\frac{d(KL(q(\mathbf{W})||p(\mathbf{W})))}{d(\sigma_{p(W)_k}^{-2})} \\ \frac{d\mathcal{L}(q(\mathbf{W}))}{d(\sigma_{p(W)_k}^{-2})} &= -\frac{1}{2} \sum_{k=1}^M (\sigma_{q(W)_k}^2 + \mu_k^2 - \sigma_{p(W)_k}^2)\end{aligned}$$

Приравняв производную к нулю, получим:

$$-\frac{1}{2} \sum_{k=1}^M (\sigma_{q(W)_k}^2 + \mu_k^2 - \sigma_{p(W)_k}^2) = 0$$

$$(\sigma_{q(W)_k}^2 + \mu_k^2 - \sigma_{p(W)_k}^2) = 0$$

$$\sigma_{p(\mathbf{W})}^2 = \mu^2 + \sigma_{q(\mathbf{W})}^2$$

Подставив полученное выражение в $KL(q(\mathbf{W})||p(\mathbf{W}))$, получим:

$$KL(q(\mathbf{W})||p(\mathbf{W})) = \frac{1}{2} \sum_{k=1}^M \ln \left(1 + \frac{\mu_k^2}{\sigma_{q(W)_k}^2} \right)$$

Чтобы избежать неопределенности $\frac{0}{0}$ и переписать выражение в векторном виде, сделаем следующую замену переменных:

$$\boldsymbol{\mu} = \boldsymbol{\gamma} \cdot \sigma_{q(\mathbf{W})}$$

$$\nu = \ln(1 + \boldsymbol{\gamma}^2)$$

Так как обучение модели будет производиться с помощью градиентных методов, сделаем следующую замену переменных, чтобы $\sigma_q(\mathbf{W})$ была всегда положительна:

$$\sigma_q(\mathbf{W}) = \ln(1 + \exp(\rho)) = \text{Softplus}(\rho)$$

Таким образом, функция потерь будет иметь следующий вид:

$$\text{Loss}(\rho, \gamma) = -\frac{\mathcal{L}(q(\mathbf{W}))}{L} = \int N(\mathbf{W} | \boldsymbol{\mu}, \sigma_q(\mathbf{W})^2 \cdot \mathbf{I}) \cdot NLL \cdot d\mathbf{W} + \frac{KL}{L}, \text{ где:}$$

$$\boldsymbol{\mu} = \gamma \cdot \sigma_q(\mathbf{W})$$

$$\sigma_q(\mathbf{W}) = \text{Softplus}(\rho)$$

$$\nu = \ln(1 + \gamma^2)$$

$$NLL = -\frac{1}{L} \sum_{i=1}^L \ln p(\mathbf{y}_i | \mathbf{x}_i, \mathbf{W})$$

$$KL = \frac{\langle \nu, \nu \rangle}{2}$$

4 Алгоритм обучения

Задаем шаг градиентного спуска α и инициализируем параметры распределения $q(\mathbf{W})$ — $\rho \leftarrow 1$ и $\gamma \leftarrow 0$. Затем повторяем, пока не достигнем критерия остановки:

1. $\sigma \leftarrow \text{Softplus}(\rho)$
2. $\boldsymbol{\mu} \leftarrow \gamma \cdot \sigma$
3. $\hat{\mathbf{W}} \leftarrow N(0, 1)$ — сэмплируем случайные веса
4. $\hat{\mathbf{W}} \leftarrow \hat{\mathbf{W}} \cdot \sigma + \boldsymbol{\mu}$ — репараметризация
5. $nll \leftarrow -\frac{1}{L} \sum_{i=1}^L \ln p(\mathbf{y}_i | \mathbf{x}_i, \hat{\mathbf{W}})$
6. $\nu \leftarrow \ln(1 + \gamma^2)$
7. $kl \leftarrow \frac{\langle \nu, \nu \rangle}{2}$
8. $l \leftarrow nll + \frac{kl}{L}$ — считаем функцию потерь
9. $\rho \leftarrow \rho - \alpha \frac{dl}{d\rho}$
10. $\gamma \leftarrow \gamma - \alpha \frac{dl}{d\gamma}$

5 Эксперименты

Для проверки своей гипотезы я выбрал [Alzheimer's Disease Dataset](#). Данные были разбиты на тренировочную и тестовую часть в пропорции 80 на 20. В качестве архитектуры была выбрана полносвязная нейронная сеть с одним скрытым слоем и функцией активации ReLU. То есть:

$$z = \text{ReLU}(\text{matmul}(x, W_1))$$

$$y = \text{Sigmoid}(\text{matmul}(z, W_2))$$

Размерность скрытого состояния z варьировалась от 1 до 37. Для каждой размерности обучались 2 модели - классическая (без регуляризации) и байесовская. Для каждой модели производилась оценка ROC-AUC на тренировочной и тестовой выборках. На рисунках 1 и 2 представлены результаты экспериментов

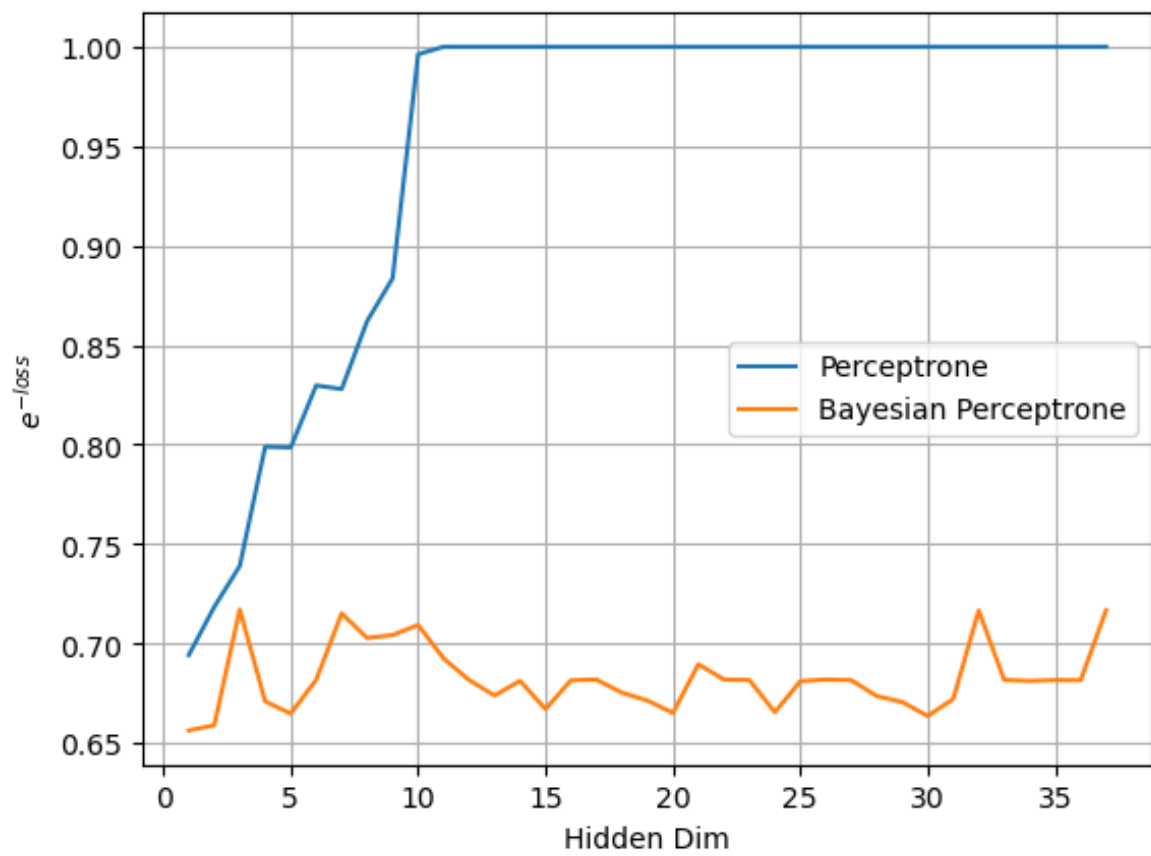


Рис. 1: Зависимость e^{-loss} от размерности скрытого состояния на тренировочных данных

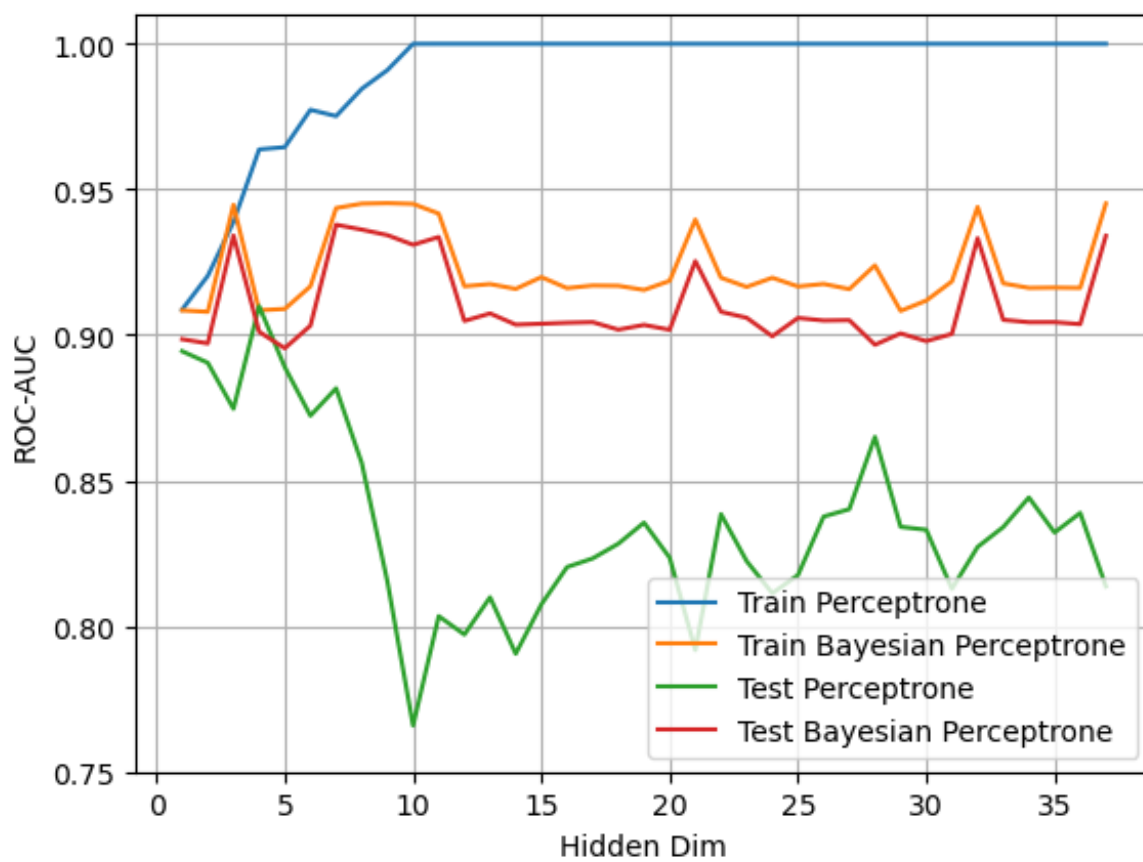


Рис. 2: Зависимость $ROC - AUC$ от размерности скрытого состояния на тренировочных и тестовых данных

6 Выводы

По результатам работы можно сделать следующие выводы:

- с ростом сложности модели байесова нейронная сеть не переобучилась;
- качество на тестовой выборке на всём рассматриваемом диапазоне гиперпараметров у байесовой нейронной сети было выше, чем у классической;
- значение ROC-AUC на тестовой выборке имеет очень высокую корреляцию со значением ELBO на тренировочной выборке (0.89 по Пирсону). Следовательно, для подбора гиперпараметров можно ориентироваться на любые монотонные преобразования от ELBO (например e^{-loss}). Это даёт нам возможность отказаться от деления на тренировочную и валидационную выборки для подбора гиперпараметров.

Так же стоит отметить, что данный подход переносится на другие архитектуры нейронных сетей (рекуррентные, свёрточные, трансформеры).

Имплементация данного подхода была выполнена с использованием PyTorch. Весь исходный код для проведения экспериментов размещён по адресу <https://github.com/dimabasow/bayesian-neural-networks>.