Эмпирические байесовские нейронные сети

Басов Дмитрий Константинович

Аннотация

Данная статья посвящена применению техники эмпирического Байеса к байесовским нейронным сетям. Концептуально идея следующая:

1. Мы используем диагональное нормальное распределение для аппроксимации апостериорного распределения весов модели — $q(W)$.
2. Априорное распределение весов модели так же задаётся диагональным распределением с нулевым матожиданием — $p(W)$.
3. Используя вариационный вывод, мы приходим к ситуации, когда ELBO зависит от $KL(q(W)||p(W))$. Так как оба распределения являются нормальными, то KL дивергенция считается аналитически.
4. Мотивация следующего этапа была взята из RVM — взять дисперсию априорного распределения весов модели $p(W)$ из данных. Там несложно берётся производная и всё получается красиво, кроме возможного деления 0/0. Но сделав замену переменных, от этой беды можно уйти.

Пункты 1–3 в принципе были описаны в статье Weight Uncertainty in Neural Networks. А вот четвёртый пункт я ни в книгах, ни в статьях не находил.

1 Обозначения и сокращения

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

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

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

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

$\mathbf{x}$ — вектор признаков

$\mathbf{y}$ — вектор целевой переменной

D — датасет — пары значений $\{\mathbf{x_i}, \mathbf{y_i}\}$, где $i = 1, \dots, L$

$\mathbf{W}$ — веса 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 весов модели (prior)

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

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

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

$q(\mathbf{W}|\boldsymbol{\theta})$ — аппроксимация апостериорного распределения весов модели

$\boldsymbol{\theta}$ — обучаемые параметры байесовской модели

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 целевой переменной $\mathbf{y}$ для неразмеченных $\mathbf{x}$, используя датасет D. Сделаем следующие преобразования:

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

В байесовском машинном обучении, веса модели являются случайной величиной. Следовательно, для того, чтобы получить предсказательное распределение $p(\mathbf{y}|\mathbf{x}, D)$, мы должны усреднить ответы, взвешенные по вероятностям возможных значений $p(\mathbf{W}|D)$, от бесконечного числа моделей.

Для аппроксимации распределения ответов модели можно воспользоваться методом Монте – Карло. То есть мы сэмплируем из распределения $p(\mathbf{W}|D)$ конечное количество весов $\hat{\mathbf{W}}_{\mathbf{1}}, \dots, \hat{\mathbf{W}}_{\mathbf{T}}$ и аппроксимируем распределение $p(\mathbf{y}|\mathbf{x}, D)$ следующим образом:

$p(\mathbf{y}|\mathbf{x}, D) \approx \dfrac{1}{T} \sum_{t=1}^{T} p(\mathbf{y}|\hat{\mathbf{W}}_{\mathbf{t}}, \mathbf{x})$, где $\hat{\mathbf{W}}_{\mathbf{t}}$ — сэмпл весов модели из $p(\mathbf{W}|D)$

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

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

Получить аналитическое решение интеграла $\int p(D|\mathbf{W}) \cdot p(\mathbf{W}) d\mathbf{W}$ 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 весов это практически невозможно. Альтернативным подходом к решению такой задачи является вариационный вывод — аппроксимация распределения $p(\mathbf{W}|D)$ распределением $q(\mathbf{W}|\boldsymbol{\theta})$, из которого сэмплировать намного проще.

3 Вариационный вывод

Идея вариационного вывода – сведение задачи байесовского вывода к задаче максимизации нижней вариационной границы (ELBO) $\mathcal{L}$, которая для распределения $q(\mathbf{W}|\boldsymbol{\theta})$ записывается следующим образом:

$$\mathcal{L}(q(\mathbf{W}|\boldsymbol{\theta})) = \int q(\mathbf{W}|\boldsymbol{\theta}) \cdot \ln \frac{p(\mathbf{D}, \mathbf{W})}{q(\mathbf{W}|\boldsymbol{\theta})} d\mathbf{W}$$

Покажем мотивацию максимизации ELBO. Для этого преобразуем выражение $\mathcal{L}(q(\mathbf{W}|\boldsymbol{\theta}))$, используя тождество $p(\mathbf{W}, D) = p(\mathbf{W}|D) \cdot p(D)$:

$$\mathcal{L}(q(\mathbf{W}|\boldsymbol{\theta})) = \int q(\mathbf{W}|\boldsymbol{\theta}) \cdot \ln \frac{p(\mathbf{D}, \mathbf{W})}{q(\mathbf{W}|\boldsymbol{\theta})} d\mathbf{W} = \int q(\mathbf{W}|\boldsymbol{\theta}) \cdot \ln \frac{p(\mathbf{W}|D) \cdot p(D)}{q(\mathbf{W}|\boldsymbol{\theta})} d\mathbf{W} =$$

$$\ln p(D) \cdot \int q(\mathbf{W}|\boldsymbol{\theta}) d\mathbf{W} - \int q(\mathbf{W}|\boldsymbol{\theta}) \cdot \ln \frac{q(\mathbf{W}|\boldsymbol{\theta})}{p(\mathbf{W}|D)} d\mathbf{W} = \ln p(D) - KL(q(\mathbf{W}|\boldsymbol{\theta})||p(\mathbf{W}|D))$$

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

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

$$\mathcal{L}(q(\mathbf{W}|\boldsymbol{\theta})) = \int q(\mathbf{W}|\boldsymbol{\theta}) \cdot \ln \frac{p(\mathbf{D}, \mathbf{W})}{q(\mathbf{W}|\boldsymbol{\theta})} d\mathbf{W} = \int q(\mathbf{W}|\boldsymbol{\theta}) \cdot \ln \frac{p(D|\mathbf{W}) \cdot p(\mathbf{W})}{q(\mathbf{W}|\boldsymbol{\theta})} d\mathbf{W} =$$

$$\int q(\mathbf{W}|\boldsymbol{\theta}) \cdot \ln p(D|\mathbf{W}) d\mathbf{W} - \int q(\mathbf{W}|\boldsymbol{\theta}) \cdot \ln \frac{q(\mathbf{W}|\boldsymbol{\theta})}{p(\mathbf{W})} =$$

$$\int q(\mathbf{W}|\boldsymbol{\theta}) \cdot \ln p(D|\mathbf{W}) d\mathbf{W} - KL(q(\mathbf{W}|\boldsymbol{\theta})||p(\mathbf{W}))$$

Таким образом, задача байесовского вывода свелась к задаче максимизации $\mathcal{L}(q(\mathbf{W}|\boldsymbol{\theta}))$ по параметрам $\boldsymbol{\theta}$:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \mathcal{L}(q(\mathbf{W}|\boldsymbol{\theta})) = \int q(\mathbf{W}|\boldsymbol{\theta}) \cdot \ln p(D|\mathbf{W}) d\mathbf{W} - KL(q(\mathbf{W}|\boldsymbol{\theta})||p(\mathbf{W}))$$

4 Задание функциональных форм распределений

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

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

$q(\mathbf{W}|\boldsymbol{\theta}) = N(\mathbf{W}|\boldsymbol{\mu}, diag(\boldsymbol{\sigma_{q(\mathbf{W})}})^2)$, где $\boldsymbol{\mu}$ и $\boldsymbol{\sigma_{q(\mathbf{W})}}$ — вектора длины M, которые вместе образуют вектор обучаемых параметров $\boldsymbol{\theta}$.

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

$$KL(q(\mathbf{W}|\boldsymbol{\theta})||p(\mathbf{W})) = \frac{1}{2}\sum_{k=1}^{M}(\frac{\sigma^2_{q(W)_k}}{\sigma^2_{p(W)_k}} + \frac{\mu_k^2}{\sigma^2_{p(W)_k}} - \ln\frac{\sigma^2_{q(W)_k}}{\sigma^2_{p(W)_k}} - 1)$$

Априорное распределение весов модели имеет нулевое математическое ожидание (из соображений симметрии), и среднеквадратическое отклонение.

В классическом байесовском выводе $\boldsymbol{\sigma_{p(\mathbf{W})}}$ должен задаваться до начала обучения, то есть являться гиперпараметром. Однако мы можем воспользоваться техникой эмпирического Байеса, то есть определить параметр априорного распределения $\boldsymbol{\sigma_{p(\mathbf{W})}}$ из данных. Для этого посчитаем $\frac{d\mathcal{L}(q(\mathbf{W}|\boldsymbol{\theta}))}{d(\sigma^{-2}_{p(W)_k})}$:

$$\frac{d\mathcal{L}(q(\mathbf{W}|\boldsymbol{\theta}))}{d(\sigma^{-2}_{p(W)_k})} = \frac{d(\int q(\mathbf{W}|\boldsymbol{\theta}) \cdot \ln p(D|\mathbf{W}) d\mathbf{W} - KL(q(\mathbf{W}|\boldsymbol{\theta})||p(\mathbf{W})))}{d(\sigma^{-2}_{p(W)_k})} = -\frac{d(KL(q(\mathbf{W}|\boldsymbol{\theta})||p(\mathbf{W})))}{d(\sigma^{-2}_{p(W)_k})} =$$

$$-\frac{1}{2}(\sigma^2_{q(W)_k} + \mu_k^2 - \sigma^2_{p(W)_k})$$

Приравняв производную к нулю, получим:

$$-\frac{1}{2}(\sigma^2_{q(W)_k} + \mu_k^2 - \sigma^2_{p(W)_k}) = 0$$

$$\sigma^2_{p(W)_k} = \sigma^2_{q(W)_k} + \mu_k^2$$

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

$$KL(q(\mathbf{W}|\boldsymbol{\theta})||p(\mathbf{W})) = \frac{1}{2}\sum_{k=1}^{M}\ln(1 + \frac{\mu_k^2}{\sigma^2_{q(W)_k}})$$

Таким образом, мы свели задачу к следующему виду:

$$\arg\max_{\boldsymbol{\mu}, \boldsymbol{\sigma_{q(\mathbf{W})}}} \mathcal{L}(q(\mathbf{W}|\boldsymbol{\mu}, \boldsymbol{\sigma_{q(\mathbf{W})}})) = \int q(\mathbf{W}|\boldsymbol{\mu}, \boldsymbol{\sigma_{q(\mathbf{W})}}) \cdot \ln p(D|\mathbf{W}) d\mathbf{W} - \frac{1}{2}\sum_{k=1}^{M}\ln(1 + \frac{\mu_k^2}{\sigma^2_{q(W)_k}})$$

5 Репараметризация

При решении задачи оптимизации при попадании в такие области, где для какого-либо веса $\sigma_{q(W)_k} = 0$ и $\mu_k = 0$, возникает неопределенность деления $\frac{0}{0}$. Чтобы избежать этой неопределенности, и чтобы $\boldsymbol{\sigma_{q(\mathbf{W})}}$ была всегда положительна, сделаем следующую замену переменных:

$\sigma_{q(W)_k} = \ln(1 + e^{\rho_k}) = Softplus(\rho_k)$

$\mu_k = \gamma_k \cdot Softplus(\rho_k)$

Тогда:

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

Таким образом, задача сводится к минимизации следующей функции потерь:

$$\underset{\boldsymbol{\rho},\boldsymbol{\gamma}}{\operatorname{argmin}}\, Loss(\boldsymbol{\rho}, \boldsymbol{\gamma}) = -\frac{\mathcal{L}(q(\mathbf{W}|\boldsymbol{\theta}))}{L} =$$
$$\int N(\mathbf{W}|\boldsymbol{\mu}, diag(\boldsymbol{\sigma_{q(\mathbf{W})}})^2) \cdot NLL \cdot d\mathbf{W} + \frac{KL}{L}$$

где:

$\boldsymbol{\sigma_{q(\mathbf{W})}} = Softplus(\boldsymbol{\rho})$

$\boldsymbol{\mu} = \boldsymbol{\gamma} \cdot Softplus(\boldsymbol{\rho})$

$NLL = -\frac{1}{L}\sum_{i=1}^{L}\ln p(\mathbf{y_i}|\mathbf{x_i}, \mathbf{W})$ — возможна аппроксимация по батчам

$KL = \frac{1}{2}\sum_{k=1}^{M}\ln(1 + \gamma_k^2)$

6 Алгоритм обучения

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

1. $\boldsymbol{\sigma} \leftarrow Softplus(\boldsymbol{\rho})$ — расчёт среднеквадратических отклонений весов
2. $\boldsymbol{\mu} \leftarrow \boldsymbol{\gamma} \odot \boldsymbol{\sigma}$ — расчёт математических ожиданий весов
3. $\hat{\mathbf{W}} \leftarrow N(0, 1)$ — сэмплирование случайных весов
4. $\hat{\mathbf{W}} \leftarrow \hat{\mathbf{W}} \odot \boldsymbol{\sigma} + \boldsymbol{\mu}$ — репараметризация
5. $nll \leftarrow -\frac{1}{L}\sum_{i=1}^{L}\ln p(\mathbf{y_i}|\mathbf{x_i}, \hat{\mathbf{W}})$ — расчёт среднего отрицательного логарифма правдоподобия
6. $kl \leftarrow \frac{1}{2}\sum_{k=1}^{M}\ln(1 + \gamma_k^2)$ — расчёт KL – дивергенции
7. $l \leftarrow nll + \frac{kl}{L}$ – расчёт функции потерь
8. $\boldsymbol{\rho} \leftarrow \boldsymbol{\rho} - \alpha\frac{dl}{d\boldsymbol{\rho}}$ — обновление $\boldsymbol{\rho}$
9. $\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma} - \alpha\frac{dl}{d\boldsymbol{\gamma}}$ — обновление $\boldsymbol{\gamma}$

7 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:

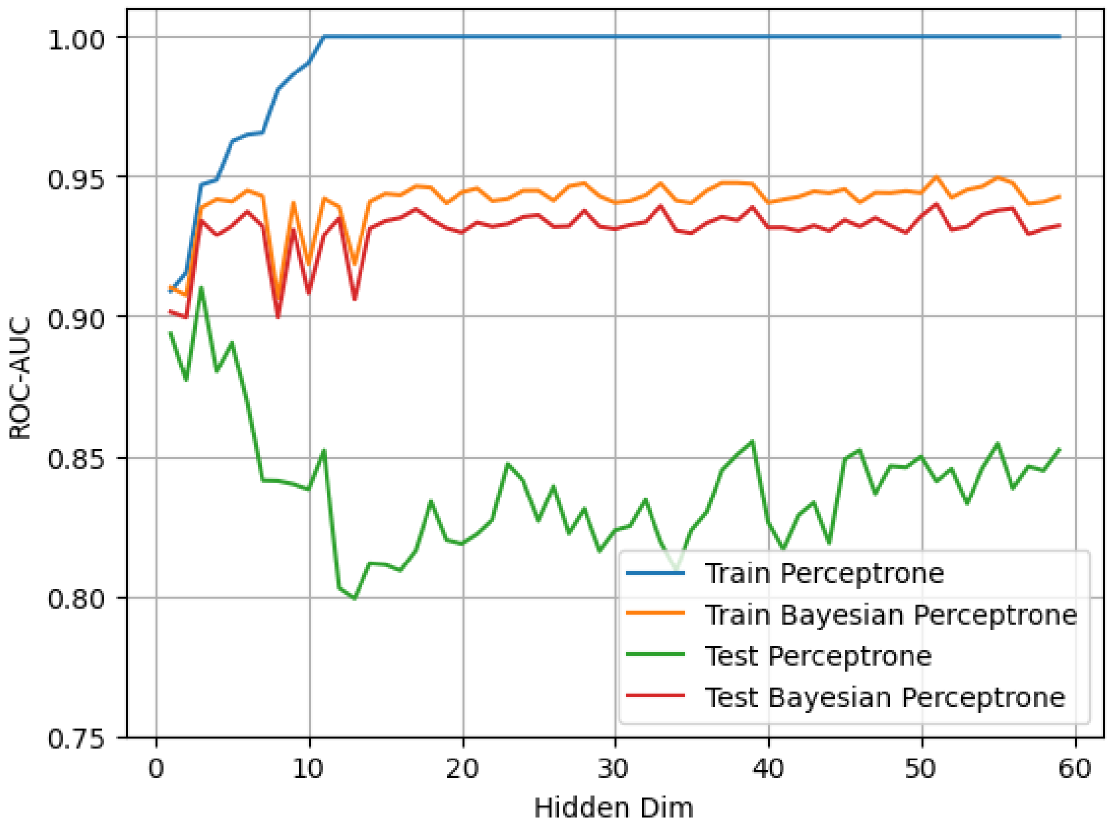

Рис. 1: Зависимость $ROC - AUC$ от размерности скрытого состояния на тренировочных и тестовых данных

$$z = ReLU(matmul(x, W_1))$$
$$y = Sigmoid(matmul(z, W_2))$$

Размерность скрытого состояния z варьировалась от 1 до 60. Для каждой размерности обучались 2 модели - классическая (без регуляризации) и байесовская. Для каждой модели производилась оценка ROC-AUC на тренировочной и тестовой выборках. На рисунке 1 представлены результаты экспериментов

8 Выводы

По результатам работы можно сделать следующие выводы:

- с ростом сложности модели байесовская нейронная сеть не переобучилась;
- значение ROC-AUC на тестовой выборке имеет очень высокую корреляцию со значением ROC-AUC на тренировочной выборке (0.97 по Пирсону). Следовательно, для подбора гиперпараметров можно ориентироваться на метрики, полученные по тренировочной выборке. 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`.