
Автоматическая регуляризации байесовских нейронных сетей

A Preprint

Басов Дмитрий Константинович

Abstract

В байесовском выводе, в отличие от гипотезы максимального правдоподобия, не делается никаких предположений о размере обучающей выборки. Это делает байесовские модели устойчивыми к переобучению.

Однако применение байесовского вывода сопряжено со следующими проблемами: заданием подходящего априорного распределения и вычислением апостериорного распределения весов модели.

В данной работе предлагается следующее решение этих проблем:

1. Применяя вариационный вывод, апостериорное распределение весов модели аппроксимируется нормальным распределением с диагональной матрицей ковариации, и задача сводится к максимизации нижней вариационной границы. В этом случае каждый вес модели определяется двумя обучаемыми параметрами.
2. Априорное распределение весов модели задается в виде нормального распределения с нулевым матожиданием и диагональной матрицей ковариации, элементы которой вычисляются из данных. Этот приём лежит в основе Relevance Vector Machine — байесовского варианта SVM.

Полученную модель можно рассматривать как ансамбль из бесконечного числа нейронных сетей, веса которых сэмпляются из нормального распределения. При этом каждый вес имеет свой индивидуальный коэффициент $L2$ регуляризации, который автоматически определяется из тренировочных данных при обучении.

1 Обозначения и сокращения

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

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

$KL(q \parallel p) = \int q(\mathbf{Z}) \ln \frac{q(\mathbf{Z})}{p(\mathbf{Z})} d\mathbf{Z}$ — 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

$p(\mathbf{y}|\mathbf{x}, D)$ — предсказательное распределение

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

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

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

$q_{\theta}(\mathbf{W})$ — аппроксимация апостериорного распределения весов модели

θ — обучаемые параметры байесовской модели

2 Введение

В классическом машинном обучении делается следующее предположение: веса модели \mathbf{W} являются пусть и неизвестной, но фиксированной величиной. В этом случае можно получить точечную оценку весов модели согласно гипотезе максимального правдоподобия.

$$\mathbf{W}_{ML} = \underset{\mathbf{W}}{\operatorname{argmax}} p(D|\mathbf{W})$$

Тогда распределение $p(\mathbf{y}|\mathbf{x}, D)$ аппроксимируется следующим образом:

$$p(\mathbf{y}|\mathbf{x}, D) \approx p(\mathbf{y}|\mathbf{x}, \mathbf{W}_{ML})$$

Однако это справедливо при условии, что количество объектов в датасете D сильно больше количества весов модели ($L \gg M$). Если это не так, веса модели \mathbf{W} могут слишком сильно подстроиться под обучающую выборку D , что чревато переобучением.

Для борьбы с переобучением используется ряд приёмов (штрафы на норму весов, early stopping, dropout), однако для их настройки требуются вычислительные ресурсы и отложенные (не участвующие в обучении) выборки данных.

Альтернативным подходом к машинному обучению является нахождение апостериорного распределения весов модели $p(\mathbf{W}|D)$ по теореме Байеса.

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

Тогда предсказательное распределение $p(\mathbf{y}|\mathbf{x}, D)$ рассчитывается следующим образом:

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

Байесовские модели можно рассматривать как ансамбль из бесконечного числа моделей, веса которых сэмпляются из распределения $p(\mathbf{W}|D)$. Такой подход устойчив к переобучению, так как о размере обучающей выборки не делается никаких предположений. Однако возникают следующие проблемы: выбор подходящего априорного распределения $p(\mathbf{W})$ и вычисление апостериорного распределения $p(\mathbf{W}|D)$.

Неудачный выбор $p(\mathbf{W})$ может сильно ухудшить качество модели, а расчёт апостериорного распределения $p(\mathbf{W}|D)$ требует вычисления интеграла по всему пространству весов модели, что для нейронных сетей практически невозможно.

В данной статье предлагается следующий подход к решению этих проблем.

1. Применяя вариационный вывод, распределение $p(\mathbf{W}|D)$ аппроксимируется распределением $q_{\theta}(\mathbf{W})$, и задача сводится к максимизации нижней вариационной границы \mathcal{L} по параметрам θ .
2. Распределение $q_{\theta}(\mathbf{W})$ задаётся в виде нормального распределения с диагональной матрицей ковариации.
3. Так как распределение $q_{\theta}(\mathbf{W})$ является нормальным, применяя трюк с репараметризацией, становится возможным использовать градиентные методы для максимизации \mathcal{L} .
4. Априорное распределение весов модели $p(\mathbf{W})$ задаётся в виде нормального распределения с нулевым матожиданием и диагональной матрицей ковариации, элементы которой определяются при обучении из датасета D . Такой подход обладает большой универсальностью, однако из-за этого теряется теоретическая устойчивость к переобучению. Идея определения некоторых параметров априорного распределения $p(\mathbf{W})$ из датасета D известна как эмпирический Байес.
5. Вводятся новые параметры γ и ρ , через которые выражаются матожидание и дисперсия распределения $q_{\theta}(\mathbf{W})$. Это нужно, чтобы избежать неопределённости деления $\frac{0}{0}$, которая может возникнуть из-за определения дисперсии распределения $p(\mathbf{W})$ из данных.

3 Постановка задачи

Задача машинного обучения с учителем в вероятностной постановке формулируется следующим образом: получить распределение вероятностей $p(\mathbf{y}|\mathbf{x}, D)$ целевой переменной \mathbf{y} для неразмеченных \mathbf{x} ,

используя информацию из датасета D . В случае параметрических моделей, которыми являются нейронные сети, информация из датасета D кодируется посредством весов модели \mathbf{W} . Сделаем следующие преобразования:

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

Пояснения:

- $p(\mathbf{y}|\mathbf{x}, D) = \int p(\mathbf{y}, \mathbf{W}|\mathbf{x}, D) d\mathbf{W}$, так как для любых случайных величин \mathbf{a} и \mathbf{b} справедливо $p(\mathbf{a}) = \int p(\mathbf{a}, \mathbf{b}) d\mathbf{b}$
- $p(\mathbf{y}, \mathbf{W}|\mathbf{x}, D) = p(\mathbf{y}|\mathbf{W}, \mathbf{x}, D) p(\mathbf{W}|\mathbf{x}, D)$, так как для любых случайных величин \mathbf{a} и \mathbf{b} справедливо $p(\mathbf{a}, \mathbf{b}) = p(\mathbf{a}|\mathbf{b}) p(\mathbf{b})$
- $p(\mathbf{y}|\mathbf{W}, \mathbf{x}, D) = p(\mathbf{y}|\mathbf{W}, \mathbf{x})$, так как вся информация из датасета D отражена в весах \mathbf{W}
- $p(\mathbf{W}|\mathbf{x}, D) = p(\mathbf{W}|D)$, так как веса модели \mathbf{W} не зависят от неразмеченных \mathbf{x} , которых не было в датасете D .

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

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

Для аппроксимации распределения ответов модели можно воспользоваться методом Монте–Карло. Идея следующая: сэмплируем конечное количество весов $\hat{\mathbf{W}}_1, \dots, \hat{\mathbf{W}}_T$ из распределения $p(\mathbf{W}|D)$ и аппроксимируем распределение $p(\mathbf{y}|\mathbf{x}, D)$ следующим образом:

$$p(\mathbf{y}|\mathbf{x}, D) \approx \frac{1}{T} \sum_{t=1}^T p(\mathbf{y}|\mathbf{x}, \hat{\mathbf{W}}_t)$$

Однако для этого нужно иметь возможность сэмплировать из распределения $p(\mathbf{W}|D)$. Получить аналитическое решение интеграла $\int p(D|\mathbf{W}) p(\mathbf{W}) d\mathbf{W}$ 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 $q_{\theta}(\mathbf{W})$, из которого сэмплировать намного проще.

4 Вариационный вывод

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

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

Покажем мотивацию максимизации ELBO. Запишем выражение для $KL(q_{\theta}(\mathbf{W}) || p(\mathbf{W}|D))$ и преобразуем его, используя тождество $p(\mathbf{W}, D) = p(\mathbf{W}|D) \cdot p(D)$:

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

Так как $\ln p(D)$ не зависит от θ , максимизация \mathcal{L} по параметрам θ ведёт к минимизации $KL(q_{\theta}(\mathbf{W}) || p(\mathbf{W}|D))$. Тем самым, при максимизации \mathcal{L} распределение $q_{\theta}(\mathbf{W})$ будет приближаться к распределению $p(\mathbf{W}|D)$.

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

$$\begin{aligned}\mathcal{L} &= \int q_{\theta}(\mathbf{W}) \cdot \ln \frac{p(\mathbf{W}, D)}{q_{\theta}(\mathbf{W})} d\mathbf{W} = \int q_{\theta}(\mathbf{W}) \cdot \ln \frac{p(D|\mathbf{W}) \cdot p(\mathbf{W})}{q_{\theta}(\mathbf{W})} d\mathbf{W} = \\ &= \int q_{\theta}(\mathbf{W}) \cdot \ln p(D|\mathbf{W}) d\mathbf{W} - \int q_{\theta}(\mathbf{W}) \cdot \ln \frac{q_{\theta}(\mathbf{W})}{p(\mathbf{W})} d\mathbf{W} = \\ &= \int q_{\theta}(\mathbf{W}) \cdot \ln p(D|\mathbf{W}) d\mathbf{W} - KL(q_{\theta}(\mathbf{W}) \parallel p(\mathbf{W})) = \\ &= \int q_{\theta}(\mathbf{W}) \cdot \sum_{i=1}^L \ln p(\mathbf{y}_i | \mathbf{x}_i, \mathbf{W}) d\mathbf{W} - KL(q_{\theta}(\mathbf{W}) \parallel p(\mathbf{W}))\end{aligned}$$

Так как в случае нейронной сети аналитически посчитать интеграл по всему пространству весов \mathbf{W} не представляется возможным, воспользуемся следующей аппроксимацией для $p(\mathbf{y}|\mathbf{x}, D)$ и \mathcal{L} :

$$\begin{aligned}p(\mathbf{y}|\mathbf{x}, D) &\approx \int p(\mathbf{y}|\mathbf{W}, \mathbf{x}) \cdot q_{\theta}(\mathbf{W}) d\mathbf{W} \approx \frac{1}{T} \sum_{t=1}^T p(\mathbf{y}|\mathbf{x}, \hat{\mathbf{W}}_t) \\ \mathcal{L} &\approx \frac{1}{S} \sum_{j=1}^S \sum_{i=1}^L \ln p(\mathbf{y}_i | \mathbf{x}_i, \hat{\mathbf{W}}_{ij}) - KL(q_{\theta}(\mathbf{W}) \parallel p(\mathbf{W}))\end{aligned}$$

где $\hat{\mathbf{W}}_t$ и $\hat{\mathbf{W}}_{ij}$ — сэмплы весов модели из распределения $q_{\theta}(\mathbf{W})$.

5 Задание функциональных форм распределений

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

$$\begin{aligned}q_{\theta}(\mathbf{W}) &= N(\mathbf{W} | \boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}_q(\mathbf{W}))^2) \\ p(\mathbf{W}) &= N(\mathbf{W} | \mathbf{0}, \text{diag}(\boldsymbol{\sigma}_p(\mathbf{W}))^2)\end{aligned}$$

Так как распределение $q_{\theta}(\mathbf{W})$ нормальное, мы можем использовать трюк с репараметризацией при сэмплировании весов, что позволяет использовать градиентные методы для оптимизации: $\hat{\mathbf{W}}_{ij} = \boldsymbol{\epsilon}_{ij} \odot \boldsymbol{\sigma}_q(\mathbf{W}) + \boldsymbol{\mu}$, где $\boldsymbol{\epsilon}_{ij}$ — сэмпл из $N(\mathbf{0}, \mathbf{I})$.

Так как распределения $p(\mathbf{W})$ и $q_{\theta}(\mathbf{W})$ нормальные, $KL(q_{\theta}(\mathbf{W}) \parallel p(\mathbf{W}))$ считается аналитически:

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

Априорное распределение весов модели $p(\mathbf{W})$ имеет нулевое математическое ожидание (из соображений симметрии), и среднеквадратическое отклонение $\boldsymbol{\sigma}_p(\mathbf{W})$. В классическом байесовском выводе параметр $\boldsymbol{\sigma}_p(\mathbf{W})$ 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.

6 Эмпирический Байес

Пусть $\boldsymbol{\alpha} = \text{diag}(\boldsymbol{\sigma}_p(\mathbf{W}))^{-2}$. Тогда выражение для $KL(q_{\theta}(\mathbf{W}) \parallel p(\mathbf{W}))$ будет иметь следующий вид:

$$KL(q_{\theta}(\mathbf{W}) \parallel p(\mathbf{W})) = \frac{1}{2} \sum_{k=1}^M (\alpha_k \cdot \sigma_{q(W)_k}^2 + \alpha_k \cdot \mu_k^2 - (\ln \sigma_{q(W)_k}^2 + \ln \alpha_k) - 1)$$

Так как в выражении \mathcal{L} интеграл $\int q_{\theta}(\mathbf{W}) \cdot \ln p(D|\mathbf{W}) d\mathbf{W}$ не зависит от параметров распределения $p(\mathbf{W})$, то:

$$\frac{\partial \mathcal{L}}{\partial \alpha_k} = -\frac{\partial(KL(q_{\theta}(\mathbf{W}) \parallel p(\mathbf{W})))}{\partial \alpha_k} = -\frac{1}{2}(\sigma_{q(W)_k}^2 + \mu_k^2 - \frac{1}{\alpha_k}) = -\frac{1}{2}(\sigma_{q(W)_k}^2 + \mu_k^2 - \sigma_{p(W)_k}^2)$$

Приравняв производную к нулю, получим:

$$\begin{aligned} -\frac{1}{2}(\sigma_{q(W)_k}^2 + \mu_k^2 - \sigma_{p(W)_k}^2) &= 0 \\ \sigma_{p(W)_k}^2 &= \sigma_{q(W)_k}^2 + \mu_k^2 \end{aligned}$$

Подставив полученное выражение для априорной дисперсии в $KL(q_{\theta}(\mathbf{W}) \parallel p(\mathbf{W}))$, получим:

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

Таким образом, мы свели задачу к максимизации \mathcal{L} по параметрам μ и $\sigma_q(\mathbf{W})$.

7 Замена переменных

При максимизации \mathcal{L} могут возникнуть ситуации, когда какой-либо вес модели перестает быть случайной величиной и вырождается в ноль ($\sigma_{q(W)_k} \rightarrow 0$ и $\mu_k \rightarrow 0$). Это приведет к неопределенности деления 0 на 0 при вычислении KL-дивергенции.

Так же при градиентной оптимизации компоненты вектора $\sigma_q(\mathbf{W})$ могут попасть в отрицательную область, что нежелательно, так как среднеквадратическое отклонение не может быть отрицательным по определению.

Чтобы избежать этих проблем, сделаем следующую замену переменных:

$$\begin{aligned} \sigma_q(\mathbf{W}) &= \ln(1 + e^{\rho}) = \text{Softplus}(\rho) \\ \mu &= \gamma \odot \sigma_q(\mathbf{W}) = \gamma \odot \text{Softplus}(\rho) \end{aligned}$$

Тогда:

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

Таким образом, задача сводится к минимизации следующей функции потерь по параметрам ρ и γ :

$$\text{loss}(\rho, \gamma) = -\frac{\mathcal{L}}{L} \approx -\frac{1}{S \cdot L} \sum_{j=1}^S \sum_{i=1}^L \ln p(\mathbf{y}_i | \mathbf{x}_i, \hat{\mathbf{W}}_{ij}) + \frac{KL}{L}$$

где:

$$\begin{aligned} \hat{\mathbf{W}}_{ij} &= \epsilon \odot \sigma + \mu \\ \epsilon &\sim N(\mathbf{0}, \mathbf{I}) \\ \sigma &= \text{Softplus}(\rho) \\ \mu &= \gamma \odot \sigma \\ KL &= \frac{1}{2} \sum_{k=1}^M \ln(1 + \gamma_k^2) \end{aligned}$$

8 Алгоритм обучения

Задаем шаг градиентного спуска α и инициализируем параметры распределения $q_{\theta}(W) = \rho$ и γ . Затем повторяем, пока не достигнем критерия останова:

1. $\sigma \leftarrow \text{Softplus}(\rho)$ — расчёт среднеквадратических отклонений весов
2. $\mu \leftarrow \gamma \odot \sigma$ — расчёт математических ожиданий весов
3. $\epsilon \leftarrow N(0, 1)$ — сэмплирование случайных весов
4. $\hat{W} \leftarrow \epsilon \odot \sigma + \mu$ — репараметризация
5. $nll \leftarrow -\frac{1}{L} \sum_{i=1}^L \ln p(\mathbf{y}_i | \mathbf{x}_i, \hat{W})$ — расчёт среднего отрицательного логарифма правдоподобия (возможна аппроксимация по батчам)
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. $\rho \leftarrow \rho - \alpha \frac{\partial l}{\partial \rho}$ — обновление ρ
9. $\gamma \leftarrow \gamma - \alpha \frac{\partial l}{\partial \gamma}$ — обновление γ

9 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 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

10 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>.

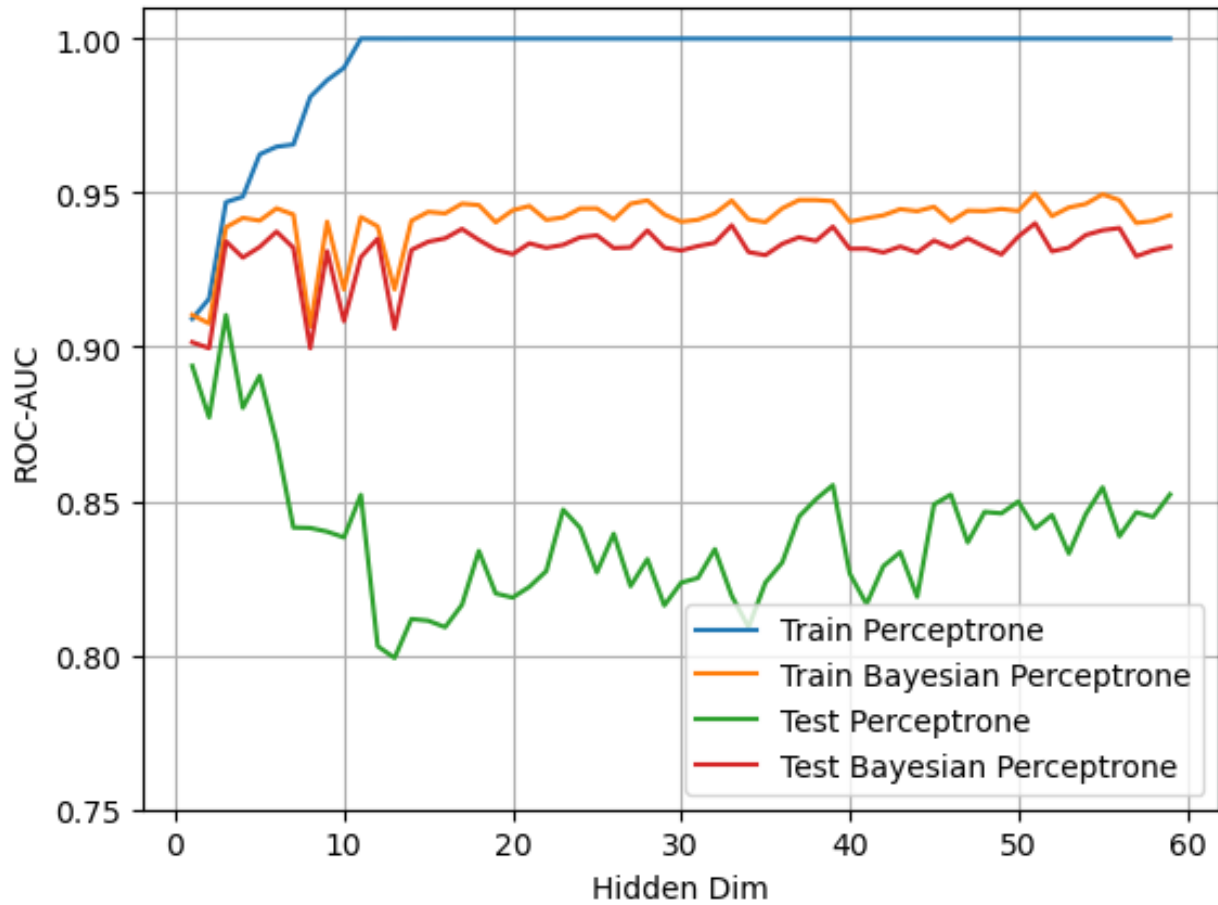


Рис. 1: Зависимость ROC-AUC от размерности скрытого состояния на тренировочных и тестовых данных