Сложность моделей глубокого обучения

Бахтеев Олег

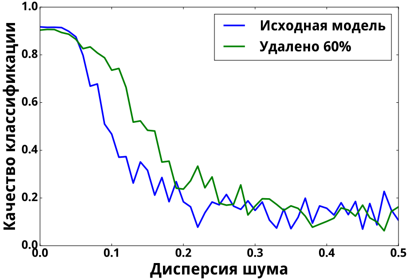
МФТИ

12.04.2017

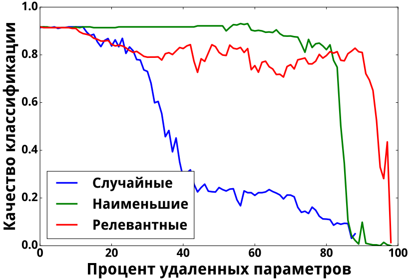
План

1. **Сложность модели**
2. **Вариационная нижняя оценка**
3. **Получение оценок для порождающих моделей**
4. **Получение оценок для разделяющих моделей**
5. **Выбор модели глубокого обучения**

Сложность модели: зачем?



Устойчивость моделей при возмущении выборки



Качество классификации при удалении параметров

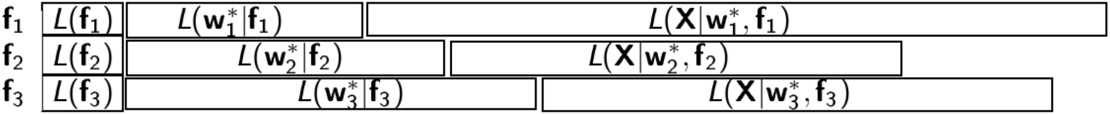
Принцип минимальной длины описания

$$\mathrm{MDL}(\mathbf{f}, \mathbf{X}) = L(\mathbf{f}) + L(\mathbf{X}|\mathbf{f}),$$

где $\mathbf{f}$ — модель, $\mathbf{X}$ — выборка, L — длина описания в битах.

$$\mathrm{MDL}(\mathbf{f}, \mathbf{X}) \sim L(\mathbf{f}) + L(\mathbf{w}^*|\mathbf{f}) + L(\mathbf{X}|\mathbf{w}^*, \mathbf{f}),$$

$\mathbf{w}^*$ — оптимальные параметры модели.



MDL и Колмогоровская сложность

Колмогоровская сложность — длина минимального кода для выборки на предварительно заданном языке.

Теорема инвариантности

Для двух сводимых по Тьюрингу языков колмогоровская сложность отличается не более чем на константу, не зависяющую от мощности выборки.

Отличия от MDL:

- Колмогоровская сложность невычислима.
- Длина кода может зависеть от выбранного языка. Для небольших выборок теорема инвариантности не дает адекватных результатов.

Оптимальная универсальная модель MDL

Пусть выборка $\mathbf{X}$ лежит в некотором конечном множестве $\mathbb{X} : \mathbf{X} \subset \mathbb{X}$.

$$\mathrm{MDL}(\mathbf{f}, \mathbf{X}) = L(\mathbf{X}|\mathbf{w}^*(\mathbf{X}), \mathbf{f}) + \mathrm{COMP}(\mathbf{f}),$$

$$L(\mathbf{X}|\mathbf{w}^*, \mathbf{f}) = -\log p(\mathbf{X}|\mathbf{w}^*(\mathbf{X}), \mathbf{f}), \quad \mathrm{COMP} = \log \sum_{\mathbf{X}' \in \mathbb{X}} P(\mathbf{X}'|\mathbf{w}^*(\mathbf{X}'), \mathbf{f}).$$

В случае, если распределение $p(\mathbf{X}|\mathbf{w})$ принадлежит экспоненциальному семейству, оценка MDL совпадает с точностью до $o(1)$ с байесовской оценкой правдоподобия (“Evidence”):

$$p(\mathbf{X}|\mathbf{f}) = \int_{\mathbf{w}} p(\mathbf{X}|\mathbf{w}) p(\mathbf{w}) d\mathbf{w},$$

где $p(\mathbf{w})$ — априорное распределение специанльного вида:

$$p(\mathbf{w}) = \frac{\sqrt{|J(\mathbf{w})|}}{\int_{\mathbf{w}'} \sqrt{|J(\mathbf{w}')|} d\mathbf{w}'},$$

$J(\mathbf{w})$ — информация Фишера.

Байесовый подход к сложности

Правдоподобие модели (“Evidence”):

$$p(\mathbf{X}|\mathbf{f}) = \int_{\mathbf{w}} p(\mathbf{X}|\mathbf{w})p(\mathbf{w}|\mathbf{f})d\mathbf{w}.$$

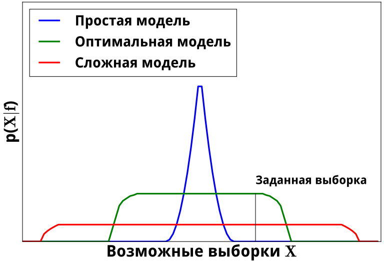
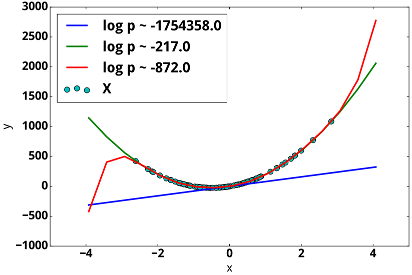


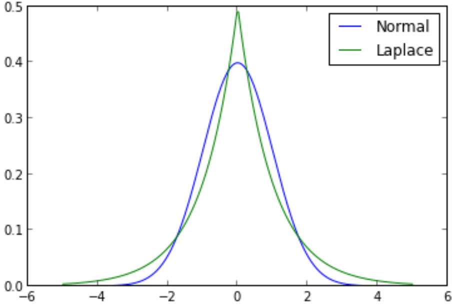
Схема выбора модели по правдоподобию



Пример: полиномы

Evidence vs MDL

Evidence	MDL
Регуляризация признаков на основе априорных знаний	-
Основывается на гипотезе о порождении выборки вне зависимости от их природы	Минимизирует длину описания выборки



Evidence vs Кросс-валидация

Оценка Evidece:

$$\log\ p(\mathbf{X}|\mathbf{f}) = \log\ p(\mathbf{x}_1|\mathbf{f}) + \log\ p(\mathbf{x}_2|\mathbf{x}_1, \mathbf{f}) + \dots + \log\ p(\mathbf{x}_n|\mathbf{x}_1, \dots, \mathbf{x}_{n-1}, \mathbf{f}).$$

Оценка leave-one-out:

$$\text{LOU} = \text{E}\log\ p(\mathbf{x}_n|\mathbf{x}_1, \dots, \mathbf{x}_{n-1}, \mathbf{f}).$$

Кросс-валидация использует среднее значение последнего члена $p(\mathbf{x}_n|\mathbf{x}_1, \dots, \mathbf{x}_{n-1}, \mathbf{f})$ для оценки сложности.

Evidence учитывает **полную** сложность описания заданной выборки, определяющую предсказательную способность модели с самого начала.

Методы получения оценок Evidence

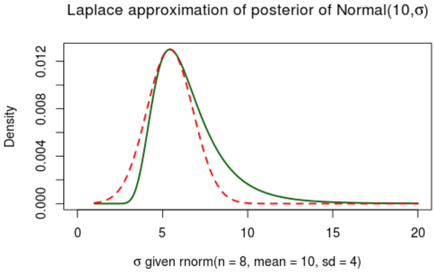
- Аппроксимация методом Лапласа

$$p(\mathbf{X}|\mathbf{f}) = \int_{\mathbf{w}} p(\mathbf{X}|\mathbf{w})p(\mathbf{w}|\mathbf{f}) = \int_{\mathbf{w}} \exp(-S(\mathbf{w})) \sim \exp S(\hat{\mathbf{w}}) \int_{\mathbf{w}} \exp(-\frac{1}{2}\Delta\mathbf{w}^{\mathsf{T}}\nabla\nabla S(\hat{\mathbf{w}})\Delta\mathbf{w}).$$

- Методы Монте-Карло

$$p(\mathbf{X}|\mathbf{f}) \sim \frac{1}{K}\sum_{\mathbf{w}\in\mathbf{W}} p(\mathbf{X}|\mathbf{w},\mathbf{f})p(\mathbf{w}|\mathbf{f}),$$

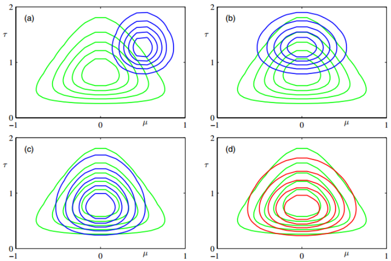
$\mathbf{W}$ — множество векторов параметров мощностью K.



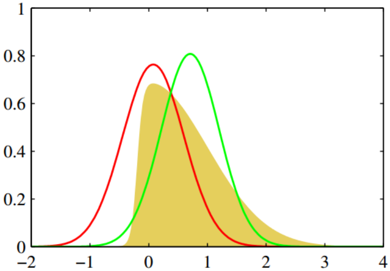
Вариационная оценка

Вариационная оценка Evidence — метод нахождения приближенного значения аналитически невычислимого распределения $p(\mathbf{w}|\mathbf{X},\mathbf{f})$ распределением $q(\mathbf{w}) \in \mathbf{Q}$. Получение вариационной нижней оценки обычно сводится к задаче минимизации

$$\mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{X})) = -\int_{\mathbf{w}} q(\mathbf{w})\log\frac{p(\mathbf{w}|\mathbf{X})}{q(\mathbf{w})}d\mathbf{w}.$$



Аппроксимация неизвестного распределения нормальным



Апроксимация Лапласа и вариационная оценка

Получение вариацонной нижней оценки

$$\log p(\mathbf{X}|\mathbf{f}) = \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{X}, \mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w} + \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{X}, \mathbf{f})) \geq$$

$$\geq \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{X}, \mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w} =$$

$$= -\mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})) + \int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathbf{X}|\mathbf{w}, \mathbf{f}) d\mathbf{w},$$

где

$$\mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})) = -\int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w}.$$

$D_{\mathbf{KL}}$

Максимизация вариационной нижней оценки

$$\int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{X}, \mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w}$$

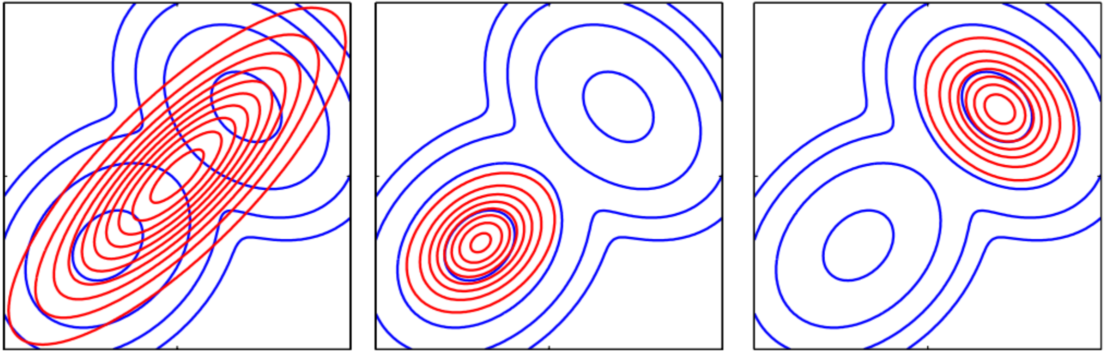
эквивалентна минимизации дивергенции между распределением распределением $q(\mathbf{w}) \in Q$ и апостериорным распределением параметров $p(\mathbf{w}|\mathbf{X}, \mathbf{f})$:

$$q = \mathrm{argmax}_{q \in Q} \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{X}, \mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w} \Leftrightarrow q = \mathrm{argmin}_{q \in Q} \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{X}, \mathbf{f})),$$

т.к.

$$\log p(\mathbf{X}|\mathbf{f}) = \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{X}, \mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w} + \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{X}, \mathbf{f})) = \mathrm{const}.$$

Пример: аппроксимация мультимодального распределения



Использование вариационной нижней оценки

Для чего используют variational inference?

- получение оценок Evidence;
- получение оценок распределений моделей со скрытыми переменными (тематическое моделирование, снижение размерности).

Зачем используют variational inference?

- сводит задачу нахождения апостериорной вероятности к методам оптимизации;
- проще масштабируется, чем аппроксимация Лапласа;
- проще в использовании, чем сэмплирующие методы.

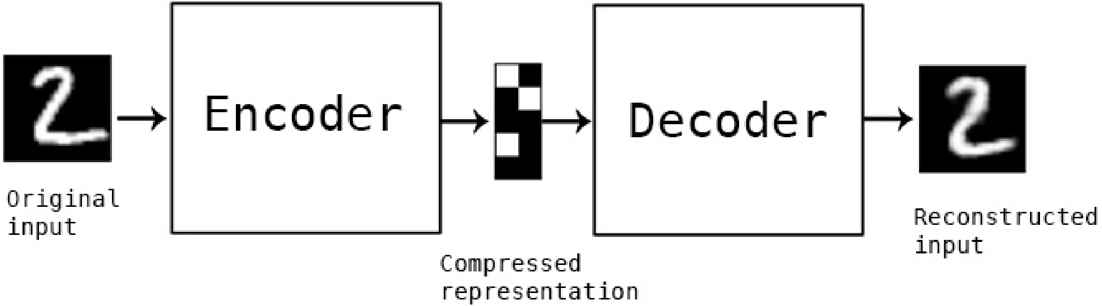
Variational Inference может давать сильно заниженную оценку.

Пример: автокодировщик

Автокодировщик — модель снижения размерности:

$$\mathbf{H} = \sigma(\mathbf{W}_e \mathbf{X}),$$

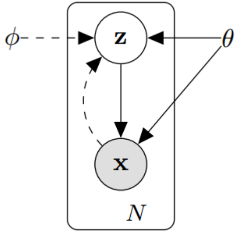
$$||\sigma(\mathbf{W}_d \mathbf{H}) - \mathbf{X}||_2^2 \to \min .$$



Вариационный автокодировщик

Пусть объекты выборки $\mathbf{X}$ порождены при условии скрытой переменной $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{h}, \mathbf{w}).$$



$p(\mathbf{h}|\mathbf{x}, \mathbf{w})$ — неизвестно.
Будем максимизировать вариационную оценку правдоподобия выборки:

$$\log p(\mathbf{x}|\mathbf{w}) \geq \mathsf{E}_{q_\phi(\mathbf{h}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{h}, \mathbf{w}) - D_{\mathrm{KL}}(q_\phi(\mathbf{h}|\mathbf{x})||p(\mathbf{h})) \to \max.$$

Распределения $q_\phi(\mathbf{h}|\mathbf{x})$ и $p(\mathbf{x}|\mathbf{h}, \mathbf{w})$ моделируются нейросетью:

$$q_\phi(\mathbf{h}|\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi(\mathbf{x})),$$

$$p(\mathbf{x}|\mathbf{h}, \mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}_w(\mathbf{h}), \boldsymbol{\sigma}^2_w(\mathbf{h})),$$

где функции $\boldsymbol{\mu}$,$\boldsymbol{\sigma}$ — выходы нейросети.

Вариационный автокодировщик: evidence

Оценка evidence получается двойным применением вариационной техники:

$$\log\, p(\mathbf{X}|\mathbf{f}) \geq \mathsf{E}_{q_{\mathbf{w}}} \log \hat{p}(\mathbf{x}|\mathbf{w}) + \log\, p(\mathbf{w}|\mathbf{f}) - \log\, q(\mathbf{w}),$$

где $q_{\mathbf{w}}$ — распределение, аппроксимирующее $p(\mathbf{w}|\mathbf{x}, \mathbf{f})$, $\log \hat{p}(\mathbf{x}|\mathbf{w})$ — вариационная оценка правдоподобия выборки.
Для оптимизации вариационных параметров применяется следующая параметризация:

$$\hat{\mathbf{w}} = \boldsymbol{\mu}_{\mathbf{w}} + \boldsymbol{\sigma}_{\mathbf{w}} \odot \boldsymbol{\epsilon}_1, \quad \hat{\mathbf{h}} = \boldsymbol{\mu}_{\mathbf{h}} + \boldsymbol{\sigma}_{\mathbf{h}}(\mathbf{h}) \odot \boldsymbol{\epsilon}_2,$$

$$\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Разделяющие модели: правдоподобие

Пусть $q \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q)$.
Тогда вариационная оценка имеет вид:

$$\int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathbf{Y}|\mathbf{X}, \mathbf{w}, \mathbf{f}) d\mathbf{w} + D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})\big) \simeq$$

$$\sum_{i=1}^{m} \log p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}_i) + D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})\big) \to \max_{\mathbf{A}_q, \boldsymbol{\mu}_q},$$

В случае, если априорное распределение параметров $p(\mathbf{w}|\mathbf{f})$ является нормальным:

$$p(\mathbf{w}|\mathbf{f}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}),$$

дивергенция $D_{\mathrm{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})$ вычисляется аналитически:

$$D_{\mathrm{KL}}\big(q(\mathbf{w}||p(\mathbf{w}|\mathbf{f})\big) = \frac{1}{2}\big(\mathrm{tr}(\mathbf{A}^{-1}\mathbf{A}_q) + (\boldsymbol{\mu} - \boldsymbol{\mu}_q)^{\mathsf{T}}\mathbf{A}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_q) - n + \ln|\mathbf{A}| - \ln|\mathbf{A}_q|\big).$$

Градиентный спуск для оценки правдоподобия

Проведем оптимизацию нейросети в режиме мультистарта из r различных начальных приближений $\mathbf{w}_1, \dots, \mathbf{w}_r$ с использованием градиентного спуска:

$$\mathbf{w}' = \mathbf{w} - \alpha \nabla \sum_{\mathbf{x} \in \mathbf{X}} \log p(\mathbf{x}, \mathbf{w}|\mathbf{f}) = \sum_{\mathbf{x} \in \mathbf{X}} \log p(\mathbf{x}|\mathbf{w}, \mathbf{f}) p(\mathbf{w}|\mathbf{f}).$$

Векторы параметров $\mathbf{w}_1, \dots, \mathbf{w}_r$ соответствуют некоторому скрытому распределению $q(\mathbf{w})$.

Энтропия

Формулу вариационной оценки можно переписать с использованием энтропии:

$$\log p(\mathbf{X}|\mathbf{f}) \geq \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{X}, \mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w} =$$

$$\mathsf{E}_{q(\mathbf{w})}[\log p(\mathbf{X}, \mathbf{w}|\mathbf{f})] - \mathsf{S}(q(\mathbf{w})),$$

где $\mathsf{S}(q(\mathbf{w}))$ — энтропия:

$$\mathsf{S}(q(\mathbf{w})) = -\int_{\mathbf{w}} q(\mathbf{w}) \log q(\mathbf{w}) d\mathbf{w}.$$

Градиентный спуск для оценки правдоподобия

При достаточно малой длине шага оптимизации α разность энтропии на различных шагах оптимизации вычисляется как:

$$\mathsf{S}(q'(\mathbf{w})) - \mathsf{S}(q(\mathbf{w})) \simeq \frac{1}{r}\sum_{g=1}^{r}\bigl(-\alpha\mathrm{Tr}[\mathbf{H}(\mathbf{w}'^{g})] - \alpha^2\mathrm{Tr}[\mathbf{H}(\mathbf{w}'^{g})\mathbf{H}(\mathbf{w}'^{g})]\bigr).$$

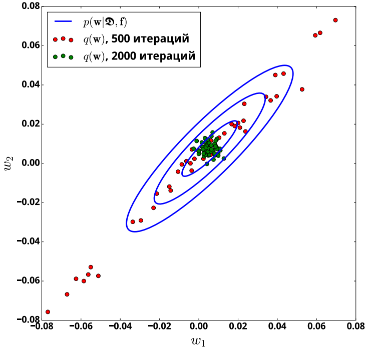
Итоговая оценка на шаге оптимизации τ:

$$\log \hat{p}(\mathbf{Y}|\mathbf{X},\mathbf{f}) \sim \frac{1}{r}\sum_{g=1}^{r} L(\mathbf{w}_{\tau}^{g},\mathbf{X},\mathbf{Y}) + \mathsf{S}(q^0(\mathbf{w})) + \frac{1}{r}\sum_{b=1}^{\tau}\sum_{g=1}^{r}\bigl(-\alpha\mathrm{Tr}[\mathbf{H}(\mathbf{w}_b^{g})] - \alpha^2\mathrm{Tr}[\mathbf{H}(\mathbf{w}_b^{g})\mathbf{H}(\mathbf{w}_b^{g})]\bigr),$$

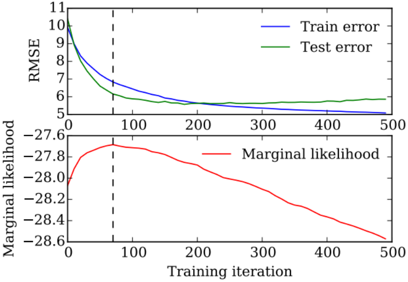
$\mathbf{w}_b^g$ — вектор параметров старта g на шаге b, $\mathsf{S}(q^0(\mathbf{w}))$ — начальная энтропия.

Переобучение

Градиентный спуск не минимизирует дивергенцию $\mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{X}))$. При приближении к моде распределения снижается оценка Evidence, что интерпретируется как переоубчение модели.



Схождение распределения к моде



Оценка начала переобучения

Стохастическая динамика Ланжевина

Модификация стохастического градиентного спуска:

$$\Delta \mathbf{w} = \alpha\nabla(\log p(\mathbf{w}) + \frac{m}{\hat{m}}\log p(\hat{\mathbf{X}}|\mathbf{w})) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \frac{\alpha}{2})$$

где $\hat{m}$ — размер подвыборки, $\hat{\mathbf{X}} \subset \mathbf{X}$ — подвыборка, шаг оптимизации α изменяется с количеством итераций:

$$\sum_{\tau=1}^{\infty} \alpha_\tau = \infty, \quad \sum_{\tau=1}^{\infty} \alpha_\tau^2 < \infty.$$

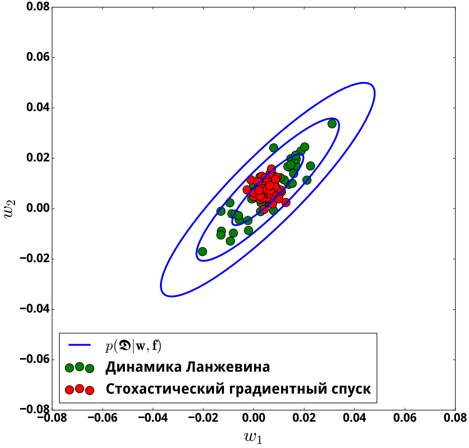
Утверждение [Welling, 2011]. Распределине $q^\tau(\mathbf{w})$ сходится к апостериорному распределению $p(\mathbf{w}|\mathbf{X}, \mathbf{f})$.

Изменение энтропии с учетом добавленного шума:

$$\hat{\mathsf{S}}(q^\tau(\mathbf{w})) \geq \frac{1}{2}|\mathbf{w}|\log(\exp(\frac{2\mathsf{S}(q^\tau(\mathbf{w}))}{|\mathbf{w}|}) + \exp(\frac{2\mathsf{S}(\epsilon)}{|\mathbf{w}|})).$$

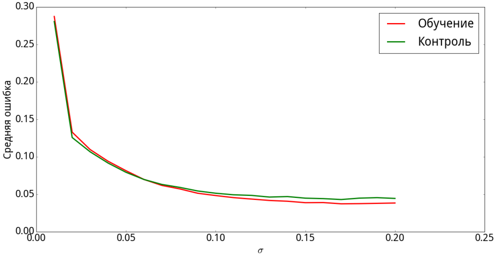
Стохастическая динамика Ланжевина

Распределения параметров после 2000 итераций:

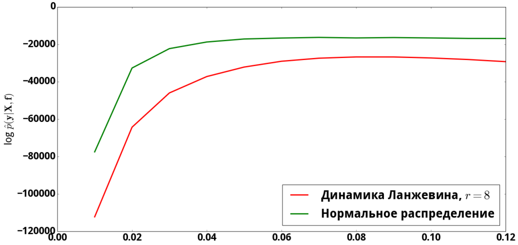


Пример: выбор константы регуляризации

Выборка MNIST, 50 нейронов на скрытом слое.



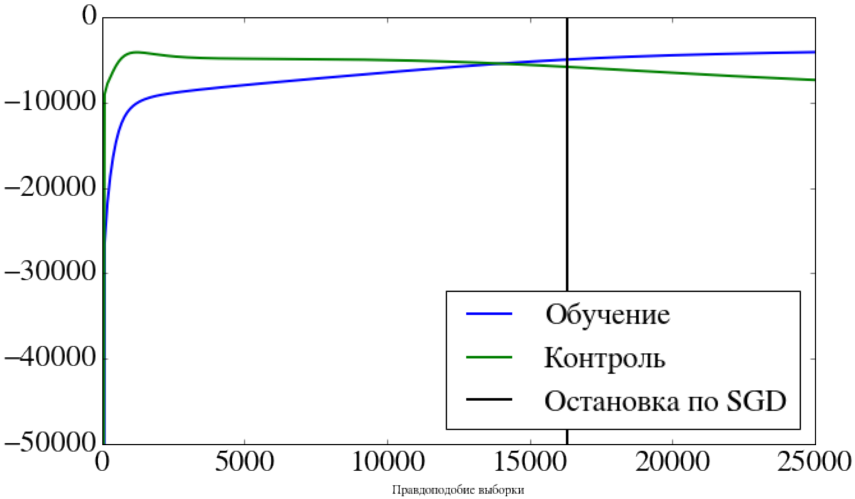
Кросс-валидация



Оценка Evidence

Пример: ранняя остановка

Выборка Boston, 50 нейронов на скрытом слое.



Выбор моделей: Graves, 2011

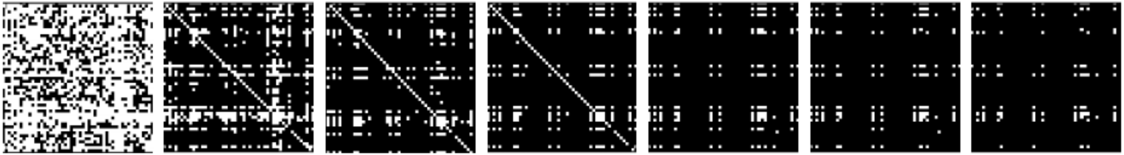
Априорное распределение: $p(\mathbf{w}|\sigma) \sim \mathcal{N}(\boldsymbol{\mu}, \sigma\mathbf{I})$.
Вариационное распределение: $q(\mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \sigma_q\mathbf{I})$.
Жадная оптимизация гиперпараметров:

$$\boldsymbol{\mu} = \hat{\mathsf{E}}\mathbf{w}, \quad \sigma = \hat{\mathsf{D}}\mathbf{w}.$$

Прунинг параметра w_i определяется относительной плотностью:

$$\frac{q(\mathbf{0})}{q(\boldsymbol{\mu}_{i,q})} = \exp(-\frac{\mu_i^2}{2\sigma_i^2}).$$

Выбор моделей: Graves, 2011



$\lambda = 0.01$ $\lambda = 0.05$ $\lambda = 0.1$ $\lambda = 0.2$ $\lambda = 0.5$ $\lambda = 1$ $\lambda = 2$

Выбор моделей: Maclaurin, Duvenaud, Adams 2015

Оптимизация гиперпараметров производится во внешнем цикле градиентными методами.

Плюсы:

1. Позволяет производить оптимизацию по произвольной дифференцируемой функции потерь.
2. Количество гиперпараметров неограничено.

Минусы:

1. Сложно реализуется технически.
2. Появляются гиперпараметры внешнего цикла оптимизации

Выбор моделей: Optimal Brain Damage

Разложим функцию потерь в окрестности точки минимума:

$$L(\mathbf{w} + \Delta \mathbf{w}) = L(\mathbf{w}) + \frac{\partial L}{\partial \mathbf{w}}^{\mathsf{T}} \Delta \mathbf{w} + 0.5 \Delta \mathbf{w}^{\mathsf{T}} \mathbf{H} \Delta \mathbf{w} + o(|\mathbf{w}|^3).$$

Выбор параметра для удаления:

$$i = \mathrm{argmin} \Delta \mathbf{w}^{\mathsf{T}} \mathbf{H} \Delta \mathbf{w}$$

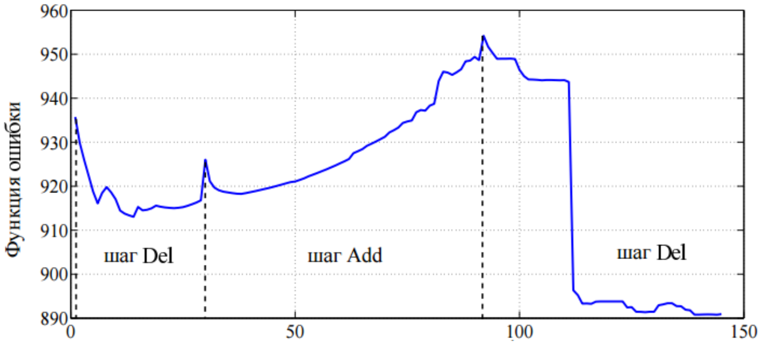
при условии:

$$\Delta \mathbf{w}_i + w_i = 0.$$

Выбор моделей: Попова, Стрижов, 2015

Критерии прореживания и наращивания сети

- Критерий оптимального прoржеивания — OBD.
- Критерий последовательного прореживания: $i = \text{argmin} L(\mathbf{f}/w_i)$.
- Критерий устойчивого прореживания — основан на анализе матрицы коварияций параметров.
- Критерий последовательного наращивания: $i = \text{argmin} L(\mathbf{f} \cup w_i)$.



Используемые материалы

1. David J. C. MacKay, Information Theory, Inference & Learning Algorithms
2. Peter Grunwald, A tutorial introduction to the minimum description length principle
3. Kuznetsov M.P., Tokmakova A.A., Strijov V.V. Analytic and stochastic methods of structure parameter estimation
4. Christopher Bishop, Pattern Recognition and Machine Learning
5. Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle, Greedy Layer-Wise Training of Deep Networks
6. Diederik P Kingma, Max Welling, Auto-Encoding Variational Bayes
7. Dougal Maclaurin, David Duvenaud, Ryan P. Adams, Early Stopping is Nonparametric Variational Inference
8. Max Welling, Yee Whye Teh, Bayesian Learning via Stochastic Gradient Langevin Dynamics

Используемые материалы

1. A. Graves, Practical Variational Inference for Neural Networks
2. D. Maclaurin, D. Duvenaud, R. P. Adams, Gradient-based Hyperparameter Optimization through Reversible Learning
3. Y. Le Cun, J. S. Denker, S. A. Solia, Optimal Brain Damage
4. М. С. Попова, В. В. Стрижов, Выбор оптимальной модели классификации физической активности по измерениям акселерометра