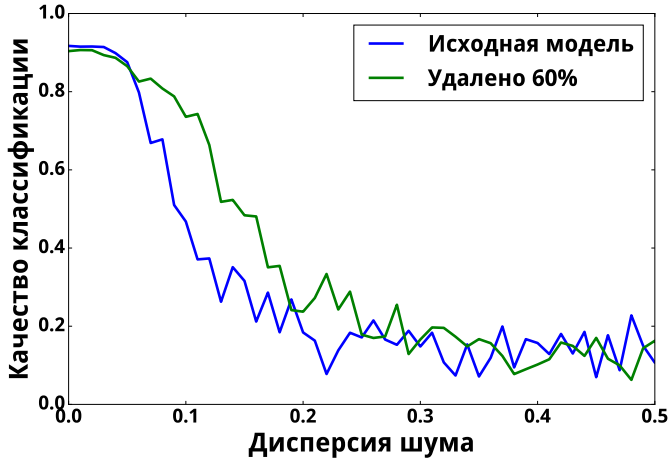
Байесовскоий выбор моделей

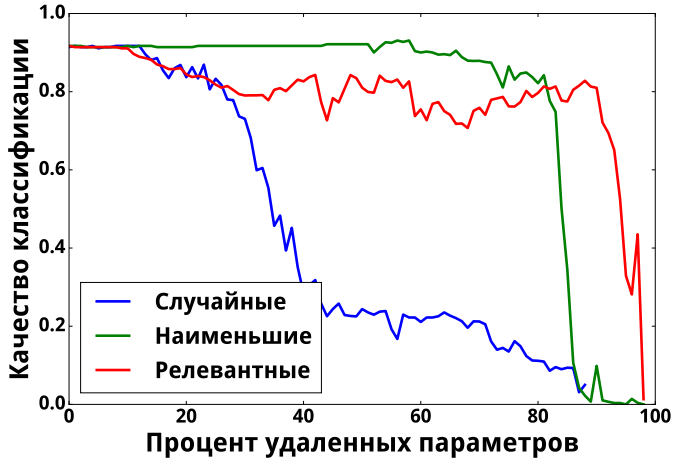
Бахтеев Олег

16.09.2019

Сложность модели: зачем?



Устойчивость моделей при возмущении выборки



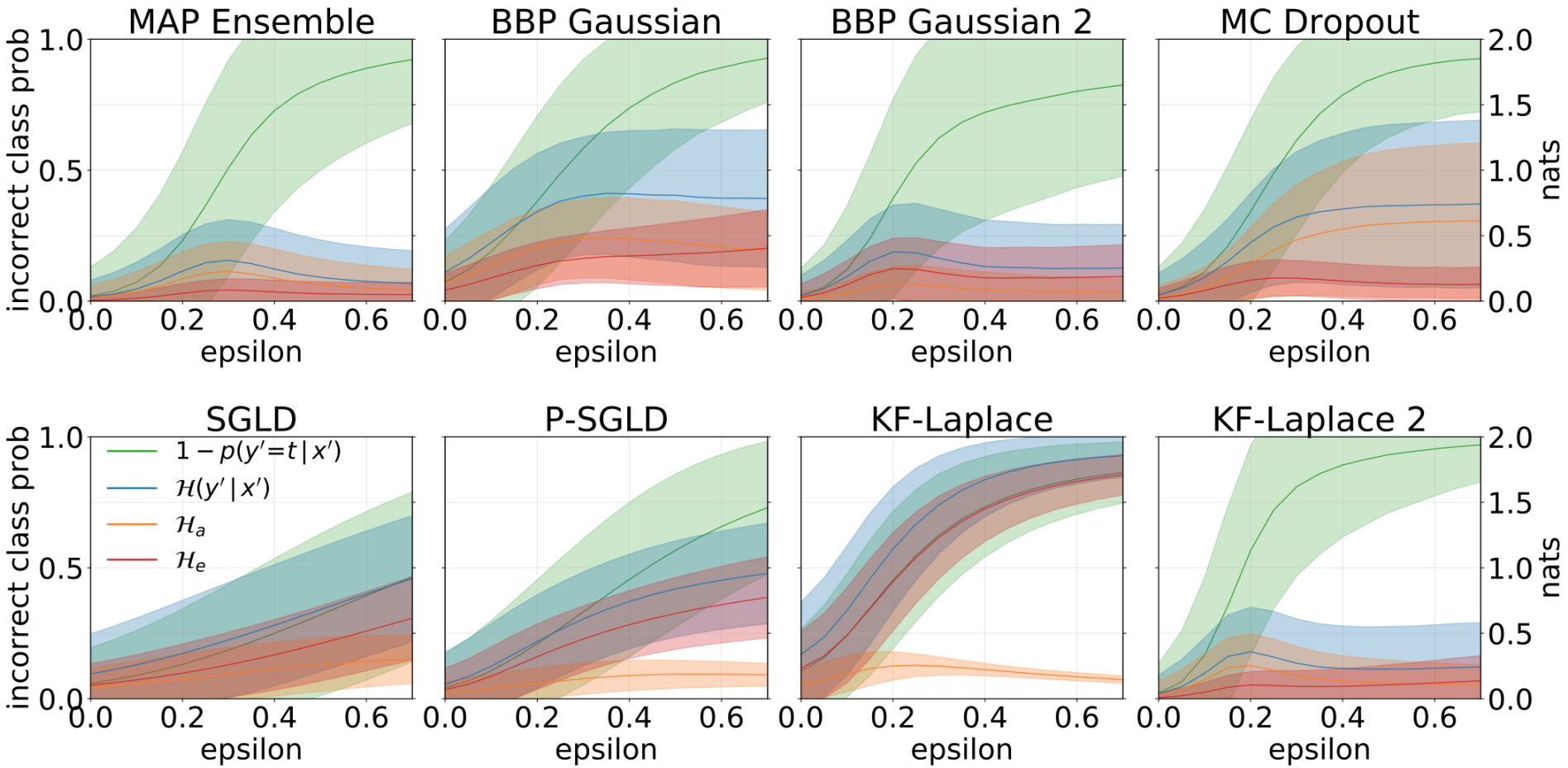
Качество классификации при удалении параметров

Сложность модели: зачем?

Model	image size	# parameters	Mult-Adds	Top 1 Acc. (%)	Top 5 Acc. (%)
Inception V2 [29]	224×224	11.2 M	1.94 B	74.8	92.2
NASNet-A (5 @ 1538)	**299×299**	**10.9 M**	**2.35 B**	**78.6**	**94.2**
Inception V3 [59]	299×299	23.8 M	5.72 B	78.0	93.9
Xception [9]	299×299	22.8 M	8.38 B	79.0	94.5
Inception ResNet V2 [57]	299×299	55.8 M	13.2 B	80.4	95.3
NASNet-A (7 @ 1920)	**299×299**	**22.6 M**	**4.93 B**	**80.8**	**95.3**
ResNeXt-101 (64 x 4d) [67]	320×320	83.6 M	31.5 B	80.9	95.6
PolyNet [68]	331×331	92 M	34.7 B	81.3	95.8
DPN-131 [8]	320×320	79.5 M	32.0 B	81.5	95.8
SENet [25]	**320×320**	**145.8 M**	**42.3 B**	**82.7**	**96.2**
NASNet-A (6 @ 4032)	**331×331**	**88.9 M**	**23.8 B**	**82.7**	**96.2**

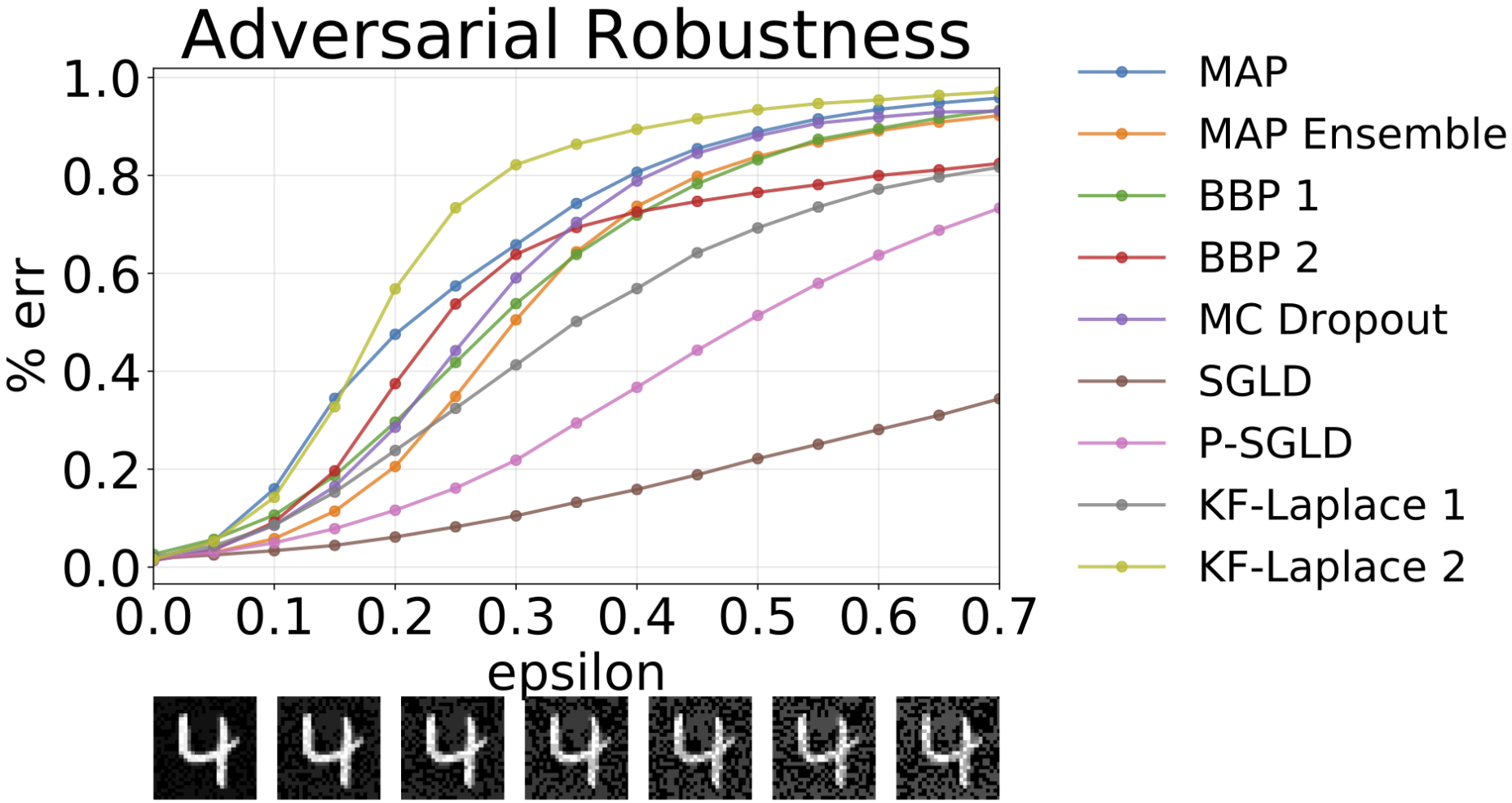
Zoph et al., 2017. Сложность моделей отличается почти в два раза при одинаковом качестве.

Устойчивость



https://github.com/JavierAntoran/Bayesian-Neural-Networks

Устойчивость



https://github.com/JavierAntoran/Bayesian-Neural-Networks

Связанный байесовский вывод

Первый уровень: выбираем оптимальные параметры:

$$\mathbf{w} = \arg\max \frac{p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})}{p(\mathfrak{D}|\mathbf{h})},$$

Второй уровень: выбираем модель, доставляющую максимум обоснованности модели.

Обоснованность модели ("Evidence"):

$$p(\mathfrak{D}|\mathbf{h}) = \int_{\mathbf{w}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})d\mathbf{w}.$$

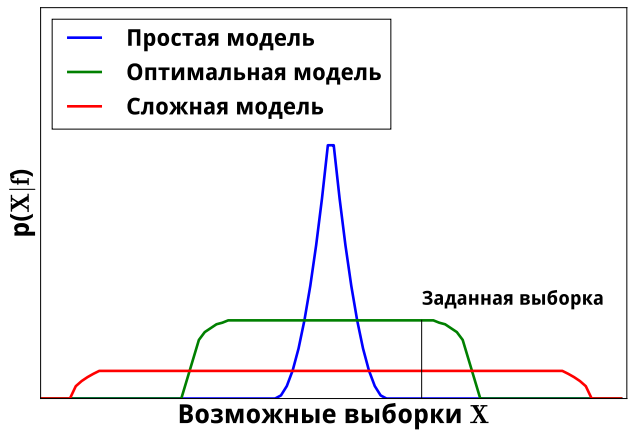
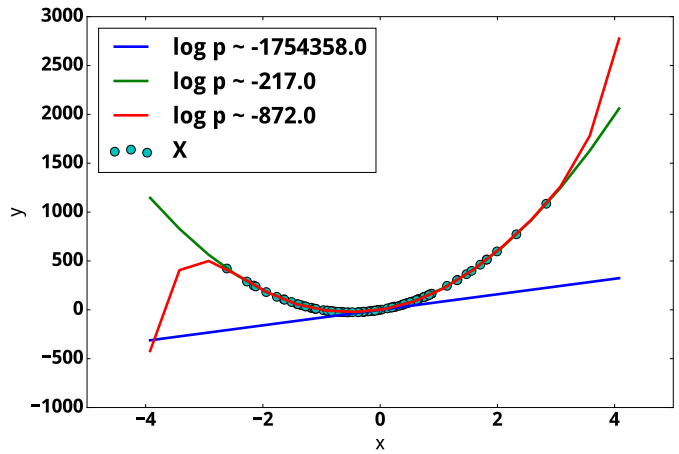


Схема выбора модели



Пример: полиномы

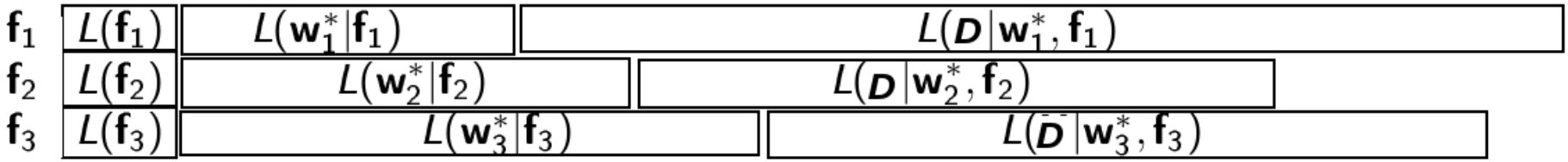
Принцип минимальной длины описания

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

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

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

$\mathbf{w}^*$ — 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 зависящую 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.

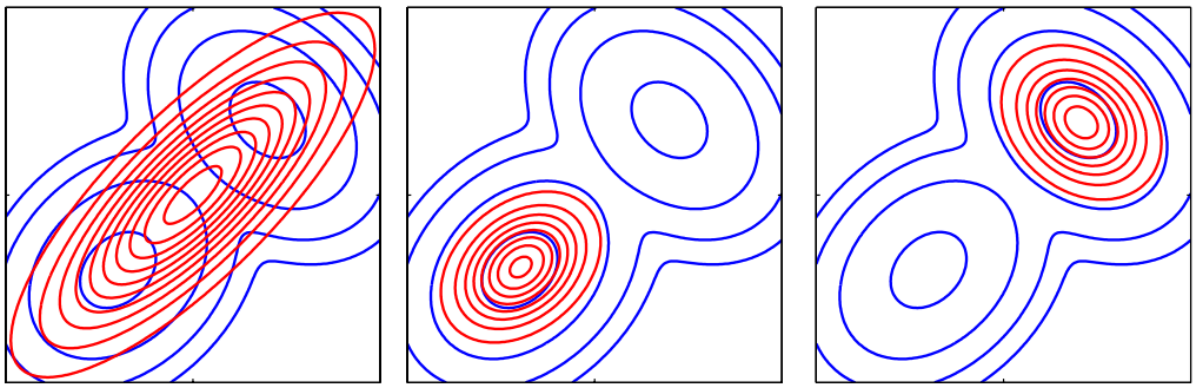
Evidence vs MDL

Evidence	**MDL**
Использует априорные знания	Независима от априорных знаний
Основывается на гипотезе о порождении выборки вне зависимости от их природы	Минимизирует длину описания выборки

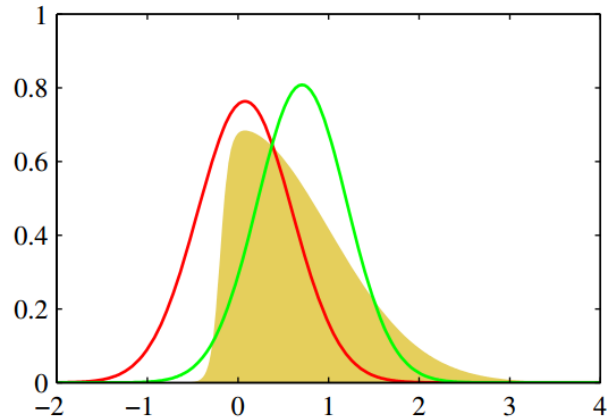
Вариационная оценка, ELBO

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

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



Вариационный вывод и expectation propogation (Bishop)



Аппроксимация Лапласа и вариационная оценка, зеленая линия (Bishop)

ELBO: нормальное распределение

“Обычная” функция потерь:

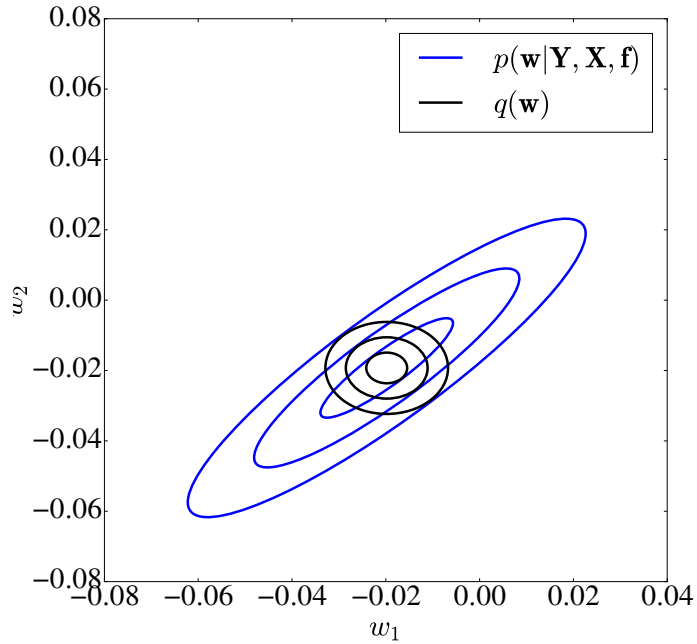
$$L = \sum_{\mathbf{x},\mathbf{y}\in\mathfrak{D}} -\log p(\mathbf{y}|\mathbf{x},\mathbf{w}) + \lambda||\mathbf{w}||_2^2.$$

Пример грубой аппроксимации нормальным диагональным распределением q

Вариационный вывод при $(p(\mathbf{w}|\mathbf{h}) \sim \mathcal{N}(\mathbf{0},\mathbf{1}))$**:**

$$L = \sum_{\mathbf{x},\mathbf{y}} \log\, p(\mathbf{y}|\mathbf{x},\hat{\mathbf{w}}) +$$

$$+\frac{1}{2}\left(\mathrm{tr}(\mathbf{A}_q) + \boldsymbol{\mu}_q^{\mathsf{T}}\mathbf{A}^{-1}\boldsymbol{\mu}_q - \ln\,|\mathbf{A}_q|\right).$$



Двухуровневая задача оптимизации

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} Q =$$

$$= \lambda^{\mathrm{Q}}_{\text{likelihood}} \mathrm{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\theta^*)} \log\, p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{h}) -$$

$$-^{\text{prior}}_{\mathrm{Q}} \mathrm{D}_{KL}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\theta^*)||p_1(\mathbf{w}|\mathbf{h})\big),$$

где

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} L = \mathrm{E}_q \log\, p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{h})$$

$$-^{\text{prior}}_{\mathrm{L}} \mathrm{D}_{KL}\big(q^*(\mathbf{w})||p_2(\mathbf{w}|\mathbf{h})\big).$$

Пример

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} Q =$$

$$= \lambda^{\mathrm{Q}}_{\text{likelihood}} \mathrm{E}_{q(\mathbf{w},\boldsymbol{\Gamma}|\theta^*)} \log\, p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{h}) -$$

$$- {}^{\text{prior}}_{\mathrm{Q}} \mathrm{D}_{KL}\big(q(\mathbf{w},\boldsymbol{\Gamma}|\theta^*)||p_1(\mathbf{w}|\mathbf{h})\big),$$

где

$$\theta^* = \arg\max_{\theta} L = \mathrm{E}_q \log\, p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{h})$$

$$- {}^{\text{prior}}_{\mathrm{L}} \mathrm{D}_{KL}\big(q^*(\mathbf{w})||p_2(\mathbf{w}|\mathbf{h})\big).$$

Что будет, если $p_1 \neq p_2$?.

- $p_1 = p_2 \sim \mathcal{N}(0, \mathbf{A}^{-1})$: задача оптимизации обоснованности. Сводится к одноуровневой.
- $p_1 \sim \mathcal{U}, p_2 \sim \mathcal{N}(0, \mathbf{A}^{-1})$: параметры доставляют максимум обоснованности, матрица $\mathbf{A}$ доставляет максимум правдоподобия.

Integrate vs maximize [MacKay]

Два подхода к назначению гиперпараметров и оптимизации параметров:

- $\mathbf{h} = \arg\max \int_{\mathbf{w}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})d\mathbf{w}$.
- $\mathbf{w} = \arg\max \int_{\mathbf{w}'} p(\mathfrak{D}|\mathbf{w}')p(\mathbf{w}')d\mathbf{w}', \quad p(\mathbf{w}') = \int_{\mathbf{h}} p(\mathbf{w}'|\mathbf{h})d\mathbf{h}$.

Informative prior vs Uninformative prior

- Informative prior: соответствует экспертным знаниям о наблюдаемой переменной
  - Пример: температура воздуха: нормальная величина с известным средним и дисперсией, соответствующими прошлым наблюдениям.
  - Соответствие апостериорного распределения априорному назовем интерпретируемостью модели.
  - Ошибка в указании информативного априорного распределения может значительно снизить итоговое качество модели.
- Uninformative prior: соответствует базовым предположениям о распределении переменной
  - Пример: температура воздуха: равномерное распределение (improper).
- Weakly-informative prior: где-то по середине
  - Пример: температура воздуха: равномерное распределение от -50 до +50.

Вопрос: $\mathbf{w} \sim \mathcal{N}(0, \mathbf{A}^{-1})$ — какой тип априорного распределения?

Пример

$$\mathbf{h}^* = \arg\max_{\mathbf{h}} Q =$$

$$= \lambda^{\mathrm{Q}}_{\text{likelihood}} \mathrm{E}_{q(\mathbf{w},\mathbf{\Gamma}|\theta^*)} \log\, p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{h}) -$$

$$-^{\text{prior}}_{\mathrm{Q}} \mathrm{D}_{KL}\big(q(\mathbf{w},\mathbf{\Gamma}|\theta^*)||p_1(\mathbf{w}|\mathbf{h})\big),$$

где

$$\theta^* = \arg\max_{\theta} L = \mathrm{E}_q \log\, p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{h})$$

$$-^{\text{prior}}_{\mathrm{L}} \mathrm{D}_{KL}\big(q^*(\mathbf{w})||p_2(\mathbf{w}|\mathbf{h})\big).$$

Гипотеза

Что будет, если $p_1 \neq p_2$?.

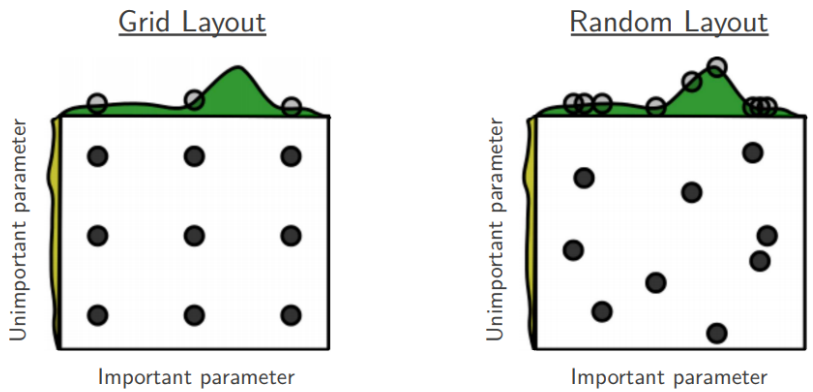
- $p_1 = p_2 =$ informative :задача оптимизации обоснованности. Сводится к одноуровневой. Может вести к переобучению.
- $p_1 =$ informative; $p_2 =$ uninformative. компромисс между информативностью и неинформативностью модели.

Базовые методы оптимизации гиперпараметров

Варианты:

- Поиск по решетке;
- Случайный поиск.

Оба метода страдают от проклятия размерности.
Случайный поиск может быть более эффективным, если пространство гиперпараметров вырождено.



Bergstra et al., 2012

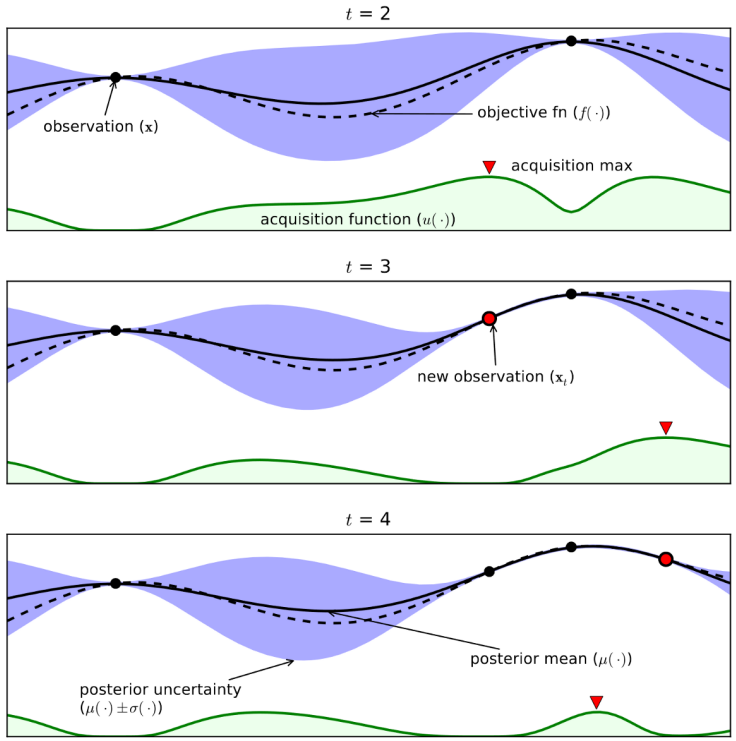
Гауссовый процесс

Идея:
Будем моделировать $Q(\theta(\mathbf{h})^*, \mathbf{h})$ гауссовым процессом, зависящим от $\mathbf{h}$.

Плюсы:

- Гибкость модели.
- Дешевле, чем обучения модели.

Минусы: кубическая сложность по количеству гиперпараметров.



Shahriari et. al, 2016. Пример работы гауссового процесса.

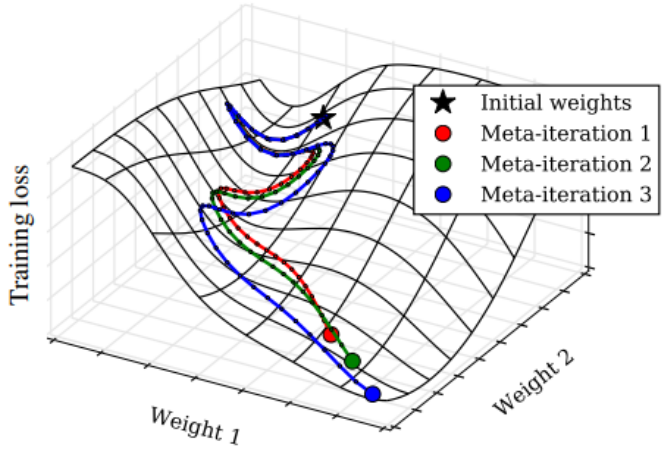
Градиентные методы

Идея: Будем производить оптимизацию вдоль всей траектории оптимизации параметров.

Плюсы:

- Оптимизация гиперпараметров будет учитывать оптимизацию параметров.
- Сложность меняется незначительно от количества гиперпараметров.

Минусы: вычилсительно дорого.



Maclaurin et. al, 2015. Пример работы.

Аналитическая формула оптимизации параметров

Утверждение (Pedregosa, 2016)

Пусть L — дифференцируемая выпуклая функция. Пусть также гессиан $\mathbf{H}^{-1}$ функции потерь L является обратимым в каждой стационарной точке. Тогда

$$\nabla_{\mathbf{h}} Q(T(\theta_0, \mathbf{h}), \mathbf{h}) = \nabla_{\mathbf{h}} Q(\theta^{\eta}, \mathbf{h}) - \nabla_{\mathbf{h}} \nabla_{\theta} L(\theta^{\eta}, \mathbf{h})^{\mathsf{T}} \mathbf{H}^{-1} \nabla_{\theta} Q(\theta^{\eta}, \mathbf{h}).$$

Жадная оптимизация гиперпараметров

На каждом шаге оптимизации параметров $\boldsymbol{\theta}$:

$$\mathbf{h}' = \mathbf{h} - \beta_{\mathbf{h}} \nabla_{\mathbf{h}} Q\big(T(\boldsymbol{\theta}, \mathbf{h}), \mathbf{h}\big) = \mathbf{h} - \beta_{\mathbf{h}} \nabla_{\mathbf{h}} Q\big(\boldsymbol{\theta} - \beta \nabla L(\boldsymbol{\theta}, \mathbf{h}), \mathbf{h}\big)\big),$$

где $\beta_{\mathbf{h}}$ — длина шага оптимизации гиперпараметров.
Метод является приближением к решению аналитической формуле в случае $\mathbf{H}^{-1} \sim \mathbf{I}$.

Параметрическая сложность

Пусть задано неинформативное априорное распределение.

Определение

Параметрической сложностью модели назовем минимальную дивергенцию между априорным и вариационным распределением:

$$C_p = \min_{\mathbf{h}} D_{\mathrm{KL}}(q||p(\mathbf{w}|\mathbf{h})).$$

Утвреждение

Устремление C_p к нулю эквивалентно снижению информативности всех параметров модели.

Вариационная оценка и эффективный размер выборки

Утверждение 2

Пусть $m \gg 0$, $\lambda > 0$, $\frac{m}{\lambda} \in \mathbb{N}$, $\frac{m}{\lambda} \gg 0$. Тогда оптимизация функции

$$\mathsf{E}_q \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - \lambda \mathsf{D}_{\mathsf{KL}}\big(q(\mathbf{w})||p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{h})\big)$$

эквивалентна оптимизации вариационной оценки обоснованности для произвольной случайной подвыборки $\hat{\mathbf{y}}, \hat{\mathbf{X}}$ мощности $\frac{m}{\lambda}$ из генеральной совокупности.

См. также [Alemi et al., 2017, Fixing Broken ELBO].

Вывод

- Чем больше коэффициент перед регуляризации модели, тем выше влияние априорного распределения.
- Интерпретация отклонения от априорного распределения различается в зависимости от его характера:
  - Неинформативное распределение: модель несет больше информации.
  - Информативное распределение: модель несет меньше информации (или априорное распределение установлено некорректно).

Список источников

- Zoph, B., Vasudevan, V., Shlens, J. and Le, Q.V., 2018. Learning transferable architectures for scalable image recognition
- David J. C. MacKay, Information Theory, Inference & Learning Algorithms
- Peter Grunwald, A tutorial introduction to the minimum description length principle
- Kuznetsov M.P., Tokmakova A.A., Strijov V.V. Analytic and stochastic methods of structure parameter estimation
- Christopher Bishop, Pattern Recognition and Machine Learning
- Diederik P Kingma, Max Welling, Auto-Encoding Variational Bayes
- Dougal Maclaurin, David Duvenaud, Ryan P. Adams, Early Stopping is Nonparametric Variational Inference
- Max Welling, Yee Whye Teh, Bayesian Learning via Stochastic Gradient Langevin Dynamics

Список источников

- A. Graves, Practical Variational Inference for Neural Networks
- Salimans, Tim, Diederik Kingma, and Max Welling, 2015. Markov chain monte carlo and variational inference: Bridging the gap
- Altieri: http://approximateinference.org/accepted/AltieriDuvenaud2015.pdf
- Stephan Mandt, Matthew D. Hoffman, David M. Blei, 2017. Stochastic Gradient Descent as Approximate Bayesian Inference
- О. Ю. Бахтеев, В. В. Стрижов, "Выбор моделей глубокого обучения субоптимальной сложности"
- А. Н. Смердов, О. Ю. Бахтеев, В. В. Стрижов, "Выбор оптимальной модели рекуррентной сети в задачах поиска парафраза"