Introduction to machine learning

Maksim Kretov

Lecture 3: Examples of methods

5vision, 2017

Course information

Course

10 lectures + 2 seminars; February-May 2017.

Schedule and up-to-date syllabus

https://goo.gl/xExEuL

Contact information and discussion

Maksim Kretov (kretovmk@gmail.com)

Slack group: https://miptmlcourse.slack.com

to get an invite, send e-mail to kretovmk@gmail.com.

Plan of the course

<u>Math and basics of ML</u>	(1-2)		*Theoretical tasks*
Some of ML methods	(3)	**Today**	
Seminar on ML basics	(4)		
Basics of neural networks	(5)	**Start playing with NNs**	*Practical tasks*
Deep learning overview	(6)		
Training deep networks	(7)		
DL for Computer Vision	(8-9)	**Solving more complex ML tasks using NNs**	
DL for time series prediction	(10-11)		
Concluding seminar	(12)		

Plan for the lecture

A. Previous lecture

B. Discriminative models

   1. Linear regression

C. Generative models

   1. Naïve Bayes classifier

D. Motivation for deep learning

E. Homework

A. Previous lecture

Machine learning tasks: Supervised learning

Given: $D = \{(\mathbf{x}_i, y_i), i = 1,..N\}$

Desired output: policy $\delta: D \to A$

Bayesian approach

Parameters θ of $p(\mathbf{x}, y|\theta)$ are treated as random variables.

Frequentist approach

Parameters θ of $p(\mathbf{x}, y|\theta)$ are unknown but fixed values.

High-level steps:

1. Select model M parameterized by θ => $p(\mathbf{x}, y|M, \theta)$
2. Infer best θ that explains given dataset D or calculate posterior distribution
3. Specify loss function $L(y, a)$
4. Design decision procedure δ

A. Previous lecture

Bayesian methods

1. Bayesian model selection: $p(m|D) = \frac{p(D|m)p(m)}{\sum_{m\epsilon M} p(D|m)p(m)}$ Select the simplest possible model

2. Infer distribution of θ conditioned on given dataset D:

$$p(\theta|D,m) = \frac{p(D|\theta,m)p(\theta|m)}{p(D|m)} = \frac{\text{likelihood} * \text{prior}}{\text{evidence}} \Rightarrow$$ Account for our uncertainty about θ

$$\Rightarrow \underline{p(\mathbf{x}, y|D)} = \int p(\mathbf{x}, y|\theta)p(\theta|\mathbf{D})d\theta$$ Hereinafter omit m for brevity

$$p(\mathbf{x}, y|D) = \underline{p(y|\mathbf{x}, D)}p(\mathbf{x}|D) = p(\mathbf{x}|y, D)p(y|D)$$ Discriminative and Generative models

3. Specify loss function $L(y, a)$. For example, accuracy $L(y,a) = \begin{cases} 0 & if \ \ y = a \\ 1 & if \ \ y \neq a \end{cases}$

4. Design decision procedure: $\delta(\mathbf{x}) = \text{argmin}_{a \in A} \mathrm{E}_{\underline{p(y|\mathbf{x},D)}}[L(y, a)]$

A. Previous lecture

Frequentist / Appr. Bayesian methods

Inferring distribution $p(\mathbf{x}, y|\theta)$ *and using it for designing policy* $\delta(\mathbf{x})$

1. Select model by cross-validation.

2. Summarize posterior distribution: MLE, MAP.

For example, MLE: $\theta_{opt} = \text{argmax}_\theta p(\boldsymbol{D}|\theta)$

$\Rightarrow \underline{p(\mathbf{x}, y|\theta_{opt})}$

3. Specify loss function $L(y, a)$. For example, accuracy $L(y, a) = \begin{cases} 0 \;\; if \;\; y = a \\ 1 \;\; if \;\; y \neq a \end{cases}$

4. Design decision procedure: $\delta(\mathbf{x}) = \text{argmin}_{a \in A} \mathrm{E}_{p(y|\mathbf{x}, D)}[L(y, a\,)]$

A. Previous lecture

Frequentist method: *Empirical Risk Minimization*

1. Select model by cross-validation.
2. Select parametric function $f(\mathbf{x}|\theta)$ to be used for prediction.
3. Specify loss function $L(y, a)$. For example: MSE.
4. Select optimal parameters θ by minimizing empirical risk w.r.t. θ:

$$\theta_{opt} = \mathrm{argmin}_\theta \frac{1}{N} \sum_{n=1}^{N} L\big(y_n, f(\mathbf{x}_{\boldsymbol{n}}|\theta)\big) + \lambda P(\theta)$$

Coefficient λ is selected by cross-validation.

B.1 Discriminative models: Linear regression

Example: Linear Regression (bias ignored for simplicity)

$p(\mathbf{x}, y|\tilde{\theta}) = p(y|\mathbf{x}, \theta')p(\mathbf{x}|\pi)$

Discriminative model, not interested in inputs' distribution

$\tilde{\theta} = \pi \cup \theta' = \pi \cup \theta \cup \sigma^2$

Model specification

$p(y|\mathbf{x}, \theta') = N(y|\theta^T\mathbf{x}, \sigma^2)$ basic model

$p(y|\mathbf{x}, \theta') = N(y|\theta^T\varphi(\mathbf{x}), \sigma^2)$ basis function expansion

$$N(y|\theta^T\mathbf{x}, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2\sigma^2}(y - \theta^T\mathbf{x})^T(y - \theta^T\mathbf{x})\right)$$

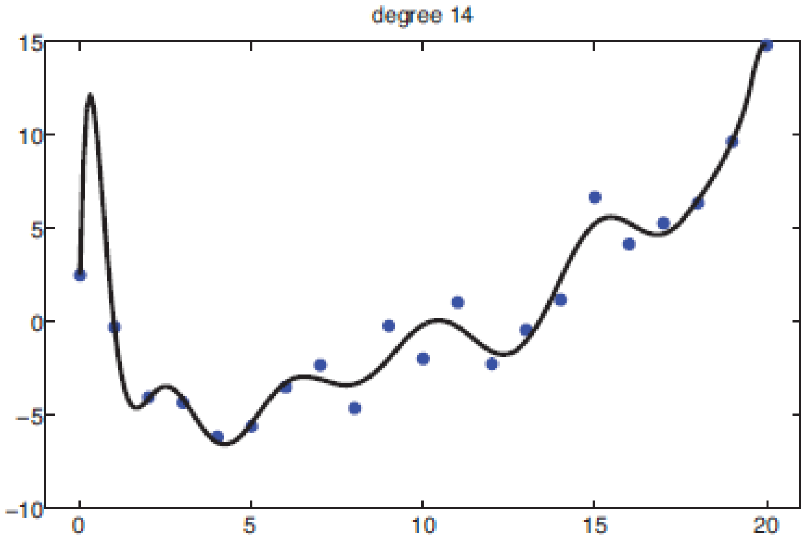
For example, polynomial basis functions: $\varphi(x) = (1, x, x^2, .. x^d)$

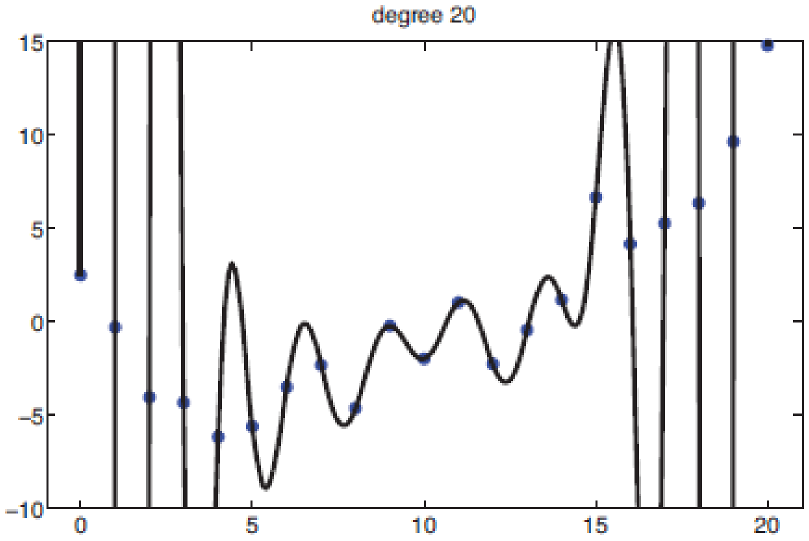
* Pictures from [3]

B.1 Discriminative models: Linear regression

Model selection

Cross-validation for models of different complexity





* Pictures from [3]

B.1 Discriminative models: Linear regression

MLE estimate of parameters of LR

Maximize likelihood of given dataset D : $\tilde{\theta}_{opt} = \text{argmax}_{\tilde{\theta}} p\left(D|\tilde{\theta}\right)$

$$\log p\left(D|\tilde{\theta}\right) = \sum_i \log p(y_i|\mathbf{x}_i, \theta') + \boxed{\sum_i \log p(\mathbf{x}_i|\pi)}$$

Will not model input distribution for discriminative model

$\text{NLL}(\theta) \triangleq -\sum_i \log p(y_i|\mathbf{x}_i, \theta')$ => task is to minimize $\text{NLL}(\theta)$

$$\text{NLL}(\theta) = C + \frac{1}{2\sigma^2}\sum_i (y_i - \theta^T\mathbf{x}_i)^T(y_i - \theta^T\mathbf{x}_i) = C + \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\theta)^T(\mathbf{y} - \mathbf{X}\theta) =$$

$$= C' + \frac{1}{2\sigma^2}\theta^T(\mathbf{X}^T\mathbf{X})\theta - \frac{1}{\sigma^2}\theta^T(\mathbf{X}^T\mathbf{y}) \Rightarrow \frac{\partial \text{NLL}(\theta)}{\partial\theta} = \frac{1}{\sigma^2}(\mathbf{X}^T\mathbf{X})\theta - \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y}$$

Setting gradient w.r.t. θ to zero: $\theta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

B.1 Discriminative models: Linear regression

Bayesian decision theory for continuous parameters

Quadratic loss: $L(y, a) = (y - a)^2$

$\rho(a|\mathbf{x}) \triangleq \mathrm{E}_{p(y|\mathbf{x},\theta')}[L(y, a)] = \mathrm{E}_{p(y|\mathbf{x},\theta')}[y^2] - 2a\mathrm{E}_{p(y|\mathbf{x},\theta')}[y] + a^2$

Let's find optimal action:

$\frac{\partial\rho(a|\mathbf{x})}{\partial a} = -2\mathrm{E}_{p(y|\mathbf{x},\theta')}[y] + 2a \Rightarrow a = \mathrm{E}_{p(y|\mathbf{x},\theta')}[y] = \bar{y}$

Optimal actions is to take mean of prediction variable

MLE solution for linear regression:

$p(y|\mathbf{x}, \theta') = N(y|\theta^T\mathbf{x}, \sigma^2)$ => mean is the best prediction

$\theta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

***Exercise:** find formula for variance*

B.1 Discriminative models: Linear regression

Next steps

MAP estimate for θ: add prior $p(\theta) = N(\theta|0, E * \sigma_0^2)$ => Ridge regression

E – unit matrix, σ_0^2- strength of prior.

Exercise: *Find MAP solution for simplified linear regression.*

Bayesian simplified linear regression

$y = C + \varepsilon$ => task is to estimate the constant using noisy measurements

Model specification: $p(y|\theta) = N(y|\theta, \sigma_\varepsilon^2)$

=> Now need to calculate posterior distribution for the mean: $p(\theta|D)$

B.1 Discriminative models: Linear regression

Bayesian simplified linear regression: posterior

Prior for parameter θ: $p(\theta) = N(\theta_0, \sigma_0^2)$

Calculating posterior:

$$p(\theta|D) = \frac{p(\theta)}{p(\boldsymbol{D})} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \exp\left(-\frac{(y_n-\theta)^2}{2\sigma_\varepsilon^2}\right) =$$

$$= \frac{p(\theta)}{p(D)} \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \exp\left(-\frac{(y_n-\theta)^2}{2\sigma_\varepsilon^2}\right) = \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left(-\frac{\left(\theta-\bar{\theta}_N\right)^2}{2\sigma_N^2}\right)$$

$$\bar{\theta}_N = \frac{N\sigma_0^2\bar{y}_N+\sigma_\varepsilon^2\theta_0}{N\sigma_0^2+\sigma_\varepsilon^2} \xrightarrow{N\to\infty} \bar{y}_N \qquad \sigma_N^2 = \frac{\sigma_\varepsilon^2\sigma_0^2}{N\sigma_0^2+\sigma_\varepsilon^2} \xrightarrow{N\to\infty} 0 \qquad \bar{y}_N = \frac{1}{N}\sum_n y_n$$

B.1 Discriminative models: Linear regression

Bayesian simplified linear regression: prediction

Joint distribution:

$$p(y|D) = \int p(y|\theta)p(\theta|D)d\theta =$$

$$= \frac{1}{2\pi\sigma_N\sigma_\varepsilon}\int \exp\left(-\frac{(y-\theta)^2}{2\sigma_\varepsilon^2} - \frac{\left(\theta-\bar{\theta}_N\right)^2}{2\sigma_N^2}\right)d\theta =$$

$$= N(y|\bar{\theta}_N, \sigma_\varepsilon^2 + \sigma_N^2)$$

Selecting an optimal action can be done using the same reasoning as we used above for MLE.

Exercise: *Check formulas for prediction.*

B.1 Discriminative models: Linear regression

Empirical risk minimization for usual linear regression

Without regularization penalty.

Prediction function: $y = \theta^T \mathbf{x}$

Loss function: $L(y, a) = (y - a)^2$

$$\theta_{opt} = \mathrm{argmin}_\theta \frac{1}{N} \sum_{n=1}^{N} L\big(y_n, f(\mathbf{x}_{\boldsymbol{n}} | \theta)\big) = \mathrm{argmin}_\theta (\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta) \Rightarrow$$

=> Same solution as MLE estimate: $\theta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

Exercise: *Check that regularization results in the same solution as MAP estimate with Gaussian prior.*

10 minute break..

C.1 Generative models: NBC

Naïve Bayes classifier example

“Naïve” because key assumption is conditional independence of features.

Consider supervised learning setting:

$\mathbf{x}$ – binary feature vector (index j), y – classification label (index c)

$$p(\mathbf{x}, y|\tilde{\theta}) = p(y|\pi)p(\mathbf{x}|y,\theta) = p(y|\pi)\prod_j p(x_j|y,\theta_j) =$$

Generative classifier

$$= \prod_c \pi_c^{\mathrm{I}(y=c)} \prod_j \prod_c p(x_j|\theta_{jc})^{\mathrm{I}(y=c)}$$

Parameters: $\pi = \{\pi_c\}, \theta = \{\theta_{jc}\}, \tilde{\theta} = \pi \cup \theta$

C.1 Generative models: NBC

Naïve Bayes classifier example

$$p(\mathbf{x}, y|\tilde{\theta}) = \prod_c \pi_c^{\mathrm{I}(y=c)} \prod_j \prod_c p(x_j|\theta_{jc})^{\mathrm{I}(y=c)} \Rightarrow$$

$$\Rightarrow \log p(D|\tilde{\theta}) = \sum_c N_c \log \pi_c + \sum_j \sum_c \sum_{i:y_i=c} \log p(x_{ij}|\theta_{jc})$$ likelihood

Now we can use maximum likelihood estimator to obtain $\tilde{\theta}_{opt}$ and get joint distribution $p(\mathbf{x}, y|\tilde{\theta})$. After that for any new example:

$$p(y|\mathbf{x}^*, \tilde{\theta}_{opt}) = \frac{p(\mathbf{x}^*|y,\tilde{\theta}_{opt})p(y|\tilde{\theta}_{opt})}{p(\mathbf{x}^*)} = \frac{p(\mathbf{x}^*|y,\tilde{\theta}_{opt})p(y|\tilde{\theta}_{opt})}{\sum_y p(\mathbf{x}^*|y,\tilde{\theta}_{opt})p(y|\tilde{\theta}_{opt})}$$ new example

C.1 Generative models: NBC

MLE estimate for parameters of NBC

Maximize likelihood of given dataset D : $\tilde{\theta}_{opt} = \mathrm{argmax}_{\tilde{\theta}} p\left(D|\tilde{\theta}\right)$

$$\log p\left(D|\tilde{\theta}\right) = \sum_{c} N_c \log \pi_c + \sum_{j} \sum_{c} \sum_{i:y_i=c} \log p\left(x_{ij}|\theta_{jc}\right)$$

Constraints: $\sum_c \pi_c = 1, \sum_c \theta_{jc} = 1 \; \forall \; j \Rightarrow$ two independent maximization tasks

Using Lagrange multipliers:

$\hat{\pi}_c = N_c / N$ (prior on classes) $\hat{\theta}_{jc} = N_{jc}/N_c$ (conditional prob. of feature)

Problem with MLE estimation: overfitting, especially if dataset is small

Exercise: *prove formulas.*

C.1 Generative models: NBC

Example: Are they Scottish?

Features = (shortbread, lager, whiskey, football). Dataset (examples X features):

English:

0	0	1	1
1	0	1	0
1	1	0	1

Scottish:

1	0	1	1
1	1	0	0
1	1	1	0

=> Following parameters:

$p(x_1 = 1|english) = 2/3$

$p(x_2 = 1|english) = 1/3$

$p(x_3 = 1|english) = 2/3$

$p(x_4 = 1|english) = 2/3$

$p(x_1 = 1|scottish) = 1$

$p(x_2 = 1|scottish) = 2/3$

$p(x_3 = 1|scottish) = 2/3$

$p(x_4 = 1|scottish) = 1/3$

* Example from [4], simplified

C.1 Generative models: NBC

Example: Are they Scottish?

New input: $\mathbf{x}^* = (1,1,1,0)$: like shortbread, lager, whiskey and doesn't watch football

$$p(s|\mathbf{x}^*) = \frac{p(\mathbf{x}^*|s)p(s)}{p(\mathbf{x}^*)} = \frac{p(\mathbf{x}^*|s)p(s)}{p(\mathbf{x}^*|s)p(s) + p(\mathbf{x}^*|e)p(e)} =$$

$$= \frac{[1 * 2/3 * 2/3 * (1 - 1/3)] * 1/2}{[1 * 2/3 * 2/3 * (1 - 1/3)] * 1/2 + [2/3 * 1/3 * 2/3 * (1 - 2/3)] * 1/2} = \frac{6}{7}$$

We calculated $p(y|\mathbf{x}^*)$, now need to find decision function $\delta(\mathbf{x}^*)$

* Example from [4], simplified

C.1 Generative models: NBC

Example: Are they Scottish?

$\mathbf{x}^* = (1,1,1,0)$

$p(s|\mathbf{x}^*) = \frac{6}{7} \quad p(e|\mathbf{x}^*) = \frac{1}{7}$

Select loss function: accuracy $L(y,a) = \begin{cases} 0 \;\; if \;\; y = a \\ 1 \;\; if \;\; y \neq a \end{cases}$

$\delta(\mathbf{x}) = \mathrm{argmin}_{a \in A} \mathrm{E}_{p(y|\mathbf{x})}[L(y, a\,)] = \mathrm{argmin}_{a \in A}[0 * p(y = a|\mathbf{x}) + 1 * p(y \neq a|\mathbf{x})]$

$= \mathrm{argmin}_{a \in A}[1 - p(y = a|\mathbf{x})] = \mathrm{argmax}_{a \in A}[p(y = a|\mathbf{x})]$

$\mathbf{Nation} = \mathrm{argmax}_a p(y = a|\mathbf{x}) = Scottish$

* Example from [4], simplified

C.1 Generative models: NBC

Bayesian NBC: motivation

Why need Bayes?

if some feature is active in all training examples, any new example $\mathbf{x}^*$ where this feature is not active does not belong to any class.

What's new in comparison with MLE?

Point estimate is replaced with full posterior distribution $p\left(\tilde{\theta}|D\right)$.

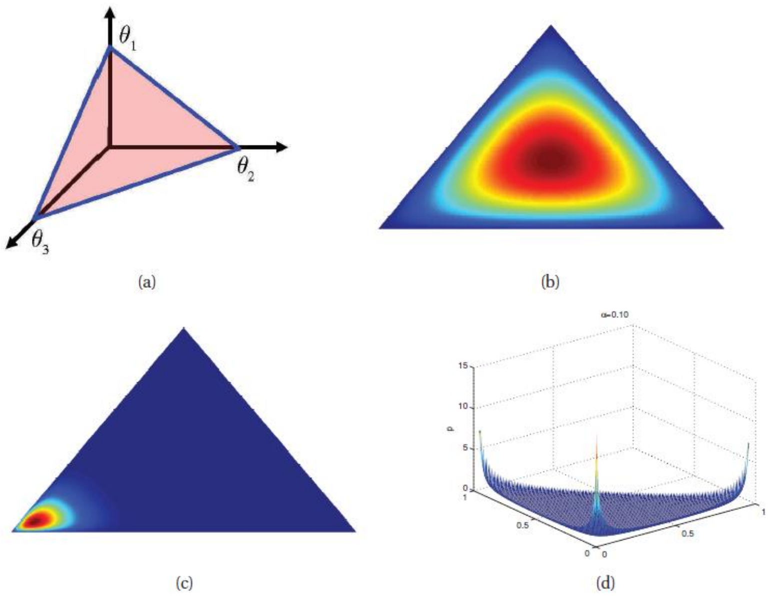
=> To be Bayesian, we need to specify prior on parameters $\tilde{\theta}$.

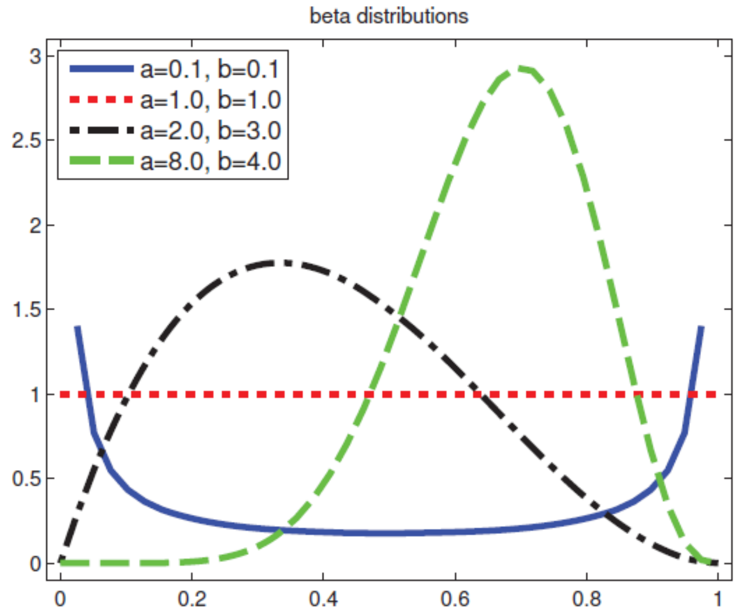
C.1 Generative models: NBC

Bayesian NBC: prior for parameters

$$p(\tilde{\theta}) = p(\pi)\prod_j \prod_c p(\theta_{jc}) = \mathrm{Dir}(\boldsymbol{\alpha} = 1)\prod_j \prod_c \mathrm{Beta}(\beta_0 = 1, \beta_1 = 1)$$

(uniform priors)





* Pictures from [3]

C.1 Generative models: NBC

Bayesian NBC: calculating posterior

For NBC can be factorized:

Prior is **conjugate** prior for the likelihood, if prior and posterior have the same form

$$p(\tilde{\theta}|D) = p(\pi|D)p(\theta|D)$$

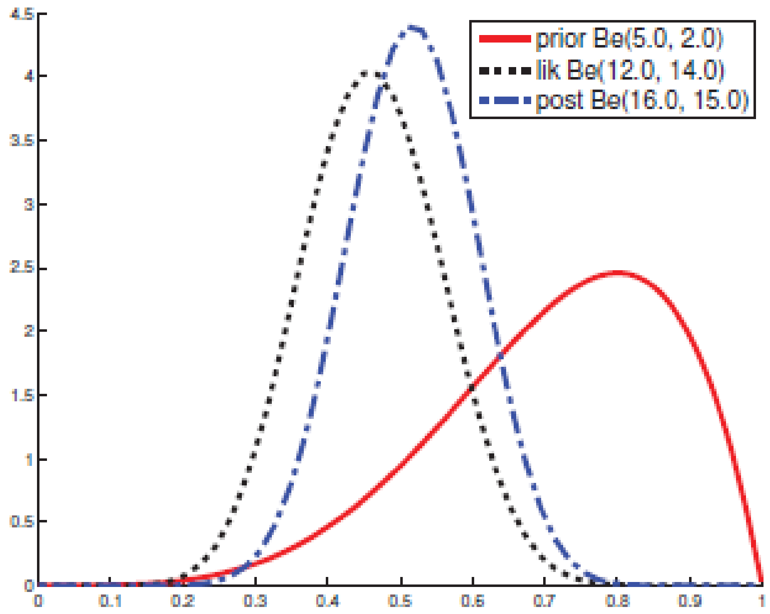
$$p(\pi|D) \sim \mathrm{Dir}(\boldsymbol{\alpha} = 1) \prod_c \pi_c^{\mathrm{I}(y=c)} \sim \mathrm{Dir}(N_1 + \alpha_1, .. N_C + \alpha_C)$$

$$p(\theta_{jc}|D) \sim \mathrm{Beta}(\beta_0 = 1, \beta_1 = 1)\theta_{jc}{}^{N_{jc}} \sim \mathrm{Beta}(\beta_0 = N_C - N_{jc} + 1, \beta_1 = N_{jc} + 1)$$

C.1 Generative models: NBC

Bayesian NBC: prior vs. posterior

Posterior for $p(\theta_{jc}|D)$



* Pictures from [3]

C.1 Generative models: NBC

Bayesian NBC: prediction

$$p(y=c|\mathbf{x}^*,D)=\frac{p(\mathbf{x}^*|y,D)p(y|D)}{p(\mathbf{x}^*|D)}\sim p(\mathbf{x}^*|y,D)p(y|D)$$

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

$$p(y|D)=\int p(y|\pi)p(\pi|D)d\pi$$

$$p(y=c|\mathbf{x}^*,D)\sim\left(\int \mathrm{Cat}(y|\pi)p(\pi|D)d\pi\right)\left(\prod_j \int \mathrm{Ber}\left(x_j^*|y,\theta_{jc}\right)p(\theta|D)d\theta_{jc}\right)$$

Answer:

$$p(y=c|\mathbf{x}^*,D)\sim\bar{\pi}_c\prod_j\left(\bar{\theta}_{jc}\right)^{\mathrm{I}(x_j=1)}\left(1-\bar{\theta}_{jc}\right)^{\mathrm{I}(x_j=0)}$$

$$\bar{\theta}_{jc}=\frac{N_{jc}+\beta_1}{N_c+\beta_0+\beta_1}\qquad\qquad \bar{\pi}_c=\frac{N_c+\alpha_c}{N+\alpha_0}$$

same as MLE estimate for $p\left(y|\mathrm{x},\tilde{\theta}\right)$, except for effect from the prior.

* Pictures from [3]

D. Motivation for deep learning

Deep Learning: machine learning algorithms based on learning multiple levels of representation / abstraction.*

Traditional methods: local smoothness assumption

Deep learning methods: complement with “compositionality” prior.

=> Just one core idea of deep learning:

Beat curse of dimensionality. Learning distributed representation can be exponentially more efficient than learning set of features that are mutually exclusive.

* Definition from NIPS 2015 Deep Learning tutorial

Next week

Concluding seminar on introduction to ML

Questions? Options:

More examples: Logistic regression, SVM

Estimators (examples of bias-variance tradeoff and regularization)

Bayesian model selection

Other?

E. Homework

For all:

1. ML: [3] ch. 3, 7 or [4] ch. 10, 13, 18 or [5] ch. 3, 7.
2. Exercises from presentation.

Refs

1. Thorough review of relevant math topics:

http://info.usherbrooke.ca/hlarochelle/ift725/review.pdf

2. Ian Goodfellow, Yoshua Bengio and Aaron Courville, Deep Learning.

3. Kevin P. Murphy, Machine Learning: A probabilistic perspective.

4. David Barber, Bayesian Reasoning and Machine Learning.

5. Sergios Theodoridis, Machine Learning: A Bayesian and optimization perspective.

Auxiliary slide #1

Bayesian linear regression: non-simplified

Joint distribution:

$$p(y|D) = \int p(y|\theta)p(\theta|D)d\theta$$

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

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

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

=> Distribution for output variable:

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