Lecture Notes and Topic Summaries

PROBABILISTIC GRAPHICAL MODELS FOR IMAGE ANALYSIS

Ondrej Skopek
Department of Computer Science
ETH Zürich
oskopek@ethz.ch

Lukas Jendele
Department of Computer Science
ETH Zürich
jendelel@ethz.ch

1 Lecture notes

1.1 Lecture 1 — Introduction to Graphical Models — 2018/09/21

Why Graphical Models?

- Neuroscience – Factor Analysis
- Image Generation – GANs
- Genomics – "Factor Analysis"
- Graph-based SLAM

1.1.1 Probabilistic Modeling

Reasoning under uncertainty – knowledge representation + automated reasoning. $\Rightarrow$ deal with uncertainty using probability.

Graphical models – provide a compact representation of two equivalent perspectives:

- Set of independencies
- Factorization of the joint dist.

Algorithms:

- Representation – semantic meaning of edges, how to use them to model tasks
- Learning – Given empirical measurements, estimate params of the model
- Inference – For a given model, use it to make decisions and reason

Machine learning: Model the task as a join probability $P(x, y), x_i \in \mathcal{X}, y_i \in \mathcal{Y}$ data + labels.

Supervised learning: given labeled training data $\mathcal{D} \subset (X \times Y)^n$ find a function $\hat{f} : X \to Y$ that correctly classifier all images, including unseen ones.

Learning a function (parametric approach). We have a class of functions $\mathcal{F} = \{f_\theta : X \to Y | \theta \in \Theta\}$. Then we can do empirical risk minimization:

$$\hat{\theta} := \arg\min_{\theta \in \Theta} \mathbb{E}_{\hat{P}} \left[\mathbb{I} \left[f_\theta(x_i) \neq y_i \right] \right].$$

The expectation is interpreted as the number of missclassifications on the training data. ERM find the function $f_{\hat{\theta}}$ that minimizes mistakes on training data.

Linear regression: $y = \theta^T x + \varepsilon$, where the MLE loss is the mean squared error: $L(\theta) = \frac{1}{n} \sum_{i=1}^{n} (\theta^T x_i - y_i)^2$

Preprint. Work in progress.

Unsupervised learning: Given a class of models $P(X;\theta), \theta in \Theta$, unlabeled data $\mathcal{D} \subset X^n$, sampled i.i.d. from $P(X)$, find the model $\hat{\theta}$ that fits the data best. I.e. want $P(X;\hat{\theta} \approx \hat{P}(X)$, or, better, $P(X;\hat{\theta} \approx P(X)$.

Likelihood: function of θ with fixed x: $L(\theta|x) := P(x|\theta)$ .. i.e. how likely is θ to have generated x?

MLE (Maximum Likelihood estimation): Given data D, let

$$L(\theta|D) := P(D|\theta) := \prod_{i=1}^{N} P(x^i|\theta)$$

. Assumes iid given θ.

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta|D)$$

. Usually minimize Negative Log Likelihood instead (NLL), because of numeric issues. Also, logarithm is monotonic, and simplifies math from products to sums.

We can show that computing MLE is the same as minimizing empirical risk. May lead to overfitting in practice, though.

1.1.2 Probabilistic Inference

Independence, conditional independence of RVs. Marginalization (expensive! exponential in number of vars!).

Factorizing joint distributions: Joint factorizes according to conditionals in Bayesian Network (BN).

Variable elimination: Marginalize and push the sums inwards, in a given order of variables. Makes a big difference computationally.

Why Bayesian networks? Because they make a lot of tasks simple and computationally efficient, and have reasonably nice interpretability.

1.2 Lecture 2 — Variational Inference — 2018/09/28

1.2.1 Expectation Maximization

Approximate inference (due to model complexity): sampling, or variational based methods. In practice, we do not observe everything in a model, and we are usually interested in unobserved (latent) variables.

Latent variable model: is $p(x, z) = p(x|z)p(z)$, where cluster membership assignment is an RV with $p(x|z = k) \sim \mathcal{N}(\mu_k, \sigma_k)$. Hence,

$$p(x) = \sum_{k=1}^{K} p(x|z = k)p(z = k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k).$$

EM algorithm: MLE with hidden variables.

General framework for partially observable data. Idea: Maximizing likelihood given "expected complete" dataset. Converges, but to local optima (because marginal likelihood increases after each EM round).

$$\log p(D) = \sum_{x \in D} \log p(x) = \sum_{x} \log(\sum_{z} p(x|z)p(z))$$

Problem: Only x is observed, but we have θ and z, both unobserved.

Algorithm:

1. Expectation: Assign values to hidden/missing variables: compute $p(z|x;\theta_t)$
2. Maximization: Maximize parameter LogLikelihood $\theta_{t+1} = \arg\max_{\theta} \mathbb{E}_{z \sim p(z|x,\theta_t)} [\log p(x, z, \theta)]$
3. Go to 1) until convergence.

1.2.2 Variational Inference

Probabilistic model: joint $p(z, x)$. Inference about unknows is through the posterior, the conditional distribution of the hidden variables given observations: $p(z|x) = p(x, z)/p(x)$. For most interesting models, the denominator $p(x)$ is intractable.

$$p(z|x, \theta) = \frac{p(z, x|\theta)}{\int_x p(z, x|\theta)dx}$$

Idea: pick family of distributions over latent variables with its own variational parameter. $q(z|\nu) = ...?$ and find variational parameters ν such that q and p are close.

Concept: Turn inference into optimization. Place a variational family of distributions over latent vars. Fit the variational parameters to be close (in KL divergence).

Convexity of functions: $f : X \to \mathbb{R}$ defined on $I = [a, b]$ is convex on I if $\forall x_1, x_2 \in I, \forall \lambda \in [0, 1]$: $f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$. f is **concave** if $-f$ is convex.

Jensen's inequality: f convex fn on interval I. If $x_1, \ldots, x_n \in I, \lambda_1, \ldots, \lambda_n \geq 0$ with $\sum_{i=1}^n \lambda_i = 1$, then

$$f(\sum_{i=1}^n \lambda_i x_i) \leq \sum_{i=1}^n \lambda_i f(x_i)$$

.

TODO: Proof by induction from convexity definition. Also $-log(x)$ is convex on $(0, \infty)$.

ELBO derivation: Let $q(z)$ be some probability distribution on z.

$$\begin{aligned}
\log p(x, \theta) &= \int q(z) \log p(x, \theta) dz \\
&= \int q(z) \log \frac{p(x, \theta) p(z|x, \theta)}{p(z|x, \theta)} dz \\
&= \int q(z) \log \frac{p(x, z, \theta)}{p(z|x, \theta)} dz \\
&= \int q(z) \log \frac{p(x, z, \theta) q(z)}{p(z|x, \theta) q(z)} dz \\
&= \int q(z) \log \frac{p(x, z, \theta)}{q(z)} dz - \int q(z) \log \frac{p(z|x, \theta)}{q(z)} dz \\
&=: \text{ELBO}(q, \theta) + \text{KL}\left[q(z) || p(z|x, \theta)\right] \\
&\geq \text{ELBO}(q, \theta)
\end{aligned}$$

By Jensen's, KL divergence is non-negative and the first term is thus a lower bound (Evidence lower bound).

TODO: What is $q(z)$?

Revisiting EM: If we can analytically calculate $q(z) := p(z|x, \theta_t)$, then

$$\begin{aligned}
\text{ELBO}(q, \theta) &= \int q(z) \log \frac{p(x, z, \theta)}{q(z)} dz \\
&= \int q(z) \log p(x, z, \theta) dz - \int q(z) \log q(z) dz \\
&= \int p(z|x, \theta_t) \log p(x, z, \theta) dz - \int p(z|x, \theta_t) \log p(z|x, \theta_t) dz \\
&= \mathcal{Q}(\theta, \theta_t) + \mathcal{H}(z|x)
\end{aligned}$$

EM maximizes ELBO instead of directly optimizing $\log p(x, \theta)$:

$$ELBO = \int ...dz = \mathbb{E}_q \left[\log p(x, z, \theta) - \log q(z)\right].$$

For e.g. Gaussian Mixture models, posterior $p(z|x;\theta_t)$ can be computed analytically.

Hence

- E-step: Compute posterior and evaluate ELBO for $q = p(z|x;\theta)$.
- M-step: $\theta_{t+1} = \arg\max_\theta \int p(z|x;\theta_t) \log p(x, z, \theta) dz$

1.2.3 Mean-field variational inference

What if we cannot find a closed form for the posterior? **We cannot use EM!**

Idea: Choose/design a variational family Q s.t. the expectations are easy to compute! (For example, they factorize).

$$q(z_1, \ldots, z_k) = \prod_{i=1}^{k} q(z_i)$$

Note: this does not contain the true posterior, as variables are most likely dependent, which this cannot capture.

Note: We can group some variables together.

ELBO for mean-field approximation:

$$\begin{aligned}
\text{ELBO}(q,\theta) &= \int q(z) \log \frac{p(x,z,\theta)}{q(z)} dz \\
&= \int \prod_i q(z_i) \log p(x, z, \theta) dz - \sum_i \int q(z_i) \log q(z_i) dz \\
&= \int q(z_j) \left(\int \prod_{i \neq j} q(z_i) \log p(x, z, \theta) \prod_{i \neq j} dz_i \right) dz_j - \int q(z_j) \log q(z_j) dz_j - \sum_{i \neq j} \int q(z_i) \log q(z_i) dz_i \\
&= \int q(z_j) \log \frac{\exp(\mathbb{E}_{i \neq j} \left[\log p(x, z, \theta)\right])}{q(z_j)} dz_j - \sum_{i \neq j} \int q(z_i) \log q(z_i) dz_i =: -\text{KL}\left[q_j || \tilde{p}_{i \neq j}\right] + \mathcal{H}(z_{i \neq j}) + c
\end{aligned}$$

Where c is a normalization constant.

1.2.4 Coordinate Ascent

KL-div is nonnegative, hence ELBO is maximal when $q(z_j) = \tilde{p}_{i \neq j} = \frac{1}{Z} \mathbb{E}_{i \neq j} \left[\log p(x, z, \theta)\right]$

Finally:

- E-step: $\forall j \text{evaluate} q^*(z_j) = \frac{1}{Z} \mathbb{E}_{i \neq j} \left[\log p(x, z, \theta)\right]$ and set $q^{t+1} = \prod_i q_i^*$
- M-step: Find $\theta_{t+1} = \arg\max_\theta ELBO(q^{t+1}, \theta)$

Overall: deterministic and fast (unlike MCMC), often works well, multiple parallel inits needed bc of local optima, ELBO is not always easy to derive.

Key idea: Bounding by convexity!

TODO: GMM with Dirichlet prior on the weights

1.3 Lecture 3 — Expectation Propagation — 2018/10/05

Key concept of VI: (cannot calculate this KL directly, because of $p(x)$, but we can see that maximizing ELBO is equiv to minimizing KL div between the posteriors, as $p(x)$ is basically a constant.

$$\begin{aligned}\text{KL}\left[q(z)||p(z|x)\right] &= \mathbb{E}_q\left[\log\frac{q(Z)}{p(Z|x)}\right]\\ &= \mathbb{E}_q\left[\log q(Z)\right] - \mathbb{E}_q\left[\log p(Z|x)\right]\\ &= \mathbb{E}_q\left[\log q(Z)\right] - \mathbb{E}_q\left[\log p(Z,x)\right] + \log p(x)\\ &= -(\mathbb{E}_q\left[\log p(Z,x)\right] - \mathbb{E}_q\left[\log q(Z)\right]) + \log p(x)\end{aligned}$$

1.3.1 Kullback-Leibler Divergence (KL-divergence)

Properties:

- $\text{KL}\left[q||p\right] \geq 0 \forall q, p$
- $\text{KL}\left[q||p\right] = 0$ iff $q = p$
- No symmetry: $\text{KL}\left[q||p\right] \neq \text{KL}\left[p||q\right]$

For $\text{KL}\left[q||p\right]$:

- If $q = p$, distributions are equal.
- If q is high, p is high – captures what we want.
- If q is high, but p is low .. **this is problematic.**
- If q is low, then expectation is 0.

Hence we use $\text{KL}\left[q||p\right]$ and not $\text{KL}\left[p||q\right]$ because we want to capture the parts where p is high, primarily.

1.3.2 Exponential families

Family of distributions over x given parameters η is the set of distributions of the form:

$$p(x|\eta) = h(x)g(\eta)\exp(\eta^T u(x))$$

or otherwise:

$$p(x|\theta) = h(x)\exp(\eta^T T(x) - A(\eta))$$

Where

- η are natural params
- $g(\eta)$ can be interpreted as a normalization
- $\mathbb{E}\left[t(x)\right] = \frac{d}{d\eta}a(\eta)$

Example: (many!) e.g. Bernoulli $p(x|\mu) = \mu^x(1-\mu)^{1-x}$ where $\eta = \frac{\mu}{1-\mu}, T(x) = x, A(\eta) = \log(1+e^\eta) = -\log(1-\mu), h(x) = 1$.

$= \exp\left(\log\left(\frac{\mu}{1-\mu}\right)x - \log(1-\mu)\right)$

Also $\mu = 1/(1+e^{-\eta})$

Conjugacy — Bayesian modeling allows to incorporate priors of family F, and likelihood of family G. F and G are conjugate if the posterior is of the same family as F (the prior).

Exponential family distributions have a conjugate prior.

For global latent vars β, local latent vars z, and observed vars x:

$$p(\beta, z, x) = p(\beta)\prod p(z_i, x_i|\beta)$$

for stochastic variational inference, assume the form of the joint of local lat. vars and observed vars given beta to be exponential, take an exponential prior on the global lat vars hence get a exp posterior.

Mean-field for conjugates

$$q(z, \beta) = q_\lambda(\beta) \prod q_{\varphi_i}(z_i)$$

Local update: $\varphi_i \leftarrow \mathbb{E}_\lambda \left[\eta_l(\beta, x_i)\right]$ for each data point

Global update: $\lambda \leftarrow \mathbb{E}_\varphi \left[\eta_g(x, z)\right]$

Alternate between these two using coordinate ascent.

1.3.3 Stochastic Variational Inference

Gradient optimization: $\lambda_{t+1} = \lambda_t + \delta \nabla_\lambda f(\lambda_t)$.

Equivalent: $\arg\max_{d\lambda} f(\lambda + d\lambda)$ s.t. $||d\lambda||^2 \leq \epsilon$.

Problem: L_2 distance not suitable for prob dists.

Natural gradient for ELBO: $\arg\max_{d\lambda} ELBO(\lambda + d\lambda)$ s.t. $D_{KL}(q_\lambda, q_{\lambda+d\lambda}) \leq \epsilon$. Where $D_{KL}(p, q) = \mathrm{KL}\left[p||q\right] + \mathrm{KL}\left[q||p\right]$

TODO slide 21

1.4 Lecture 4 — Repetition and Stochastic Variational Inference — 2018/10/12

1.5 Lecture 5 — Sequential Data — 2018/10/19

1.6 Lecture 6 — Dimensionality Reduction — 2018/10/26

1.7 Lecture 7 — Summary Dimensionality Reduction and State Space Models — 2018/11/02

1.8 Lecture 8 — Guest Lecture: Generative Adversarial Networks — 2018/11/09

Google talk. N/A for exam?

1.9 Lecture 9 — Autoencoding Variational Bayes — 2018/11/16

1.10 Lecture 10 — Score Function Estimators — 2018/11/23

1.11 Lecture 11 — Evaluating Deep Representation Learning — 2018/11/30

1.12 Lecture 12 — Guest lecture: Temporal Point Processes and Bayesian Non-parametrics — 2018/12/14

N/A for exam.

1.13 Lecture 13 — Guest lecture — 2018/12/21

N/A for exam.

2 Topics

2.1 EM Algorithm

2.1.1 Concept and examples e.g. EM for Gaussian Mixture Model

2.1.2 Why does the algorithm converge?

2.2 Variational Inference

2.2.1 Concept and examples e.g. Mean-field for Gaussian Mixture with Dirichlet Prior

2.2.2 Expectation Propagation and α divergence

2.2.3 Examples e.g. Expectation Propagation for Gaussian Mixture

2.2.4 Explain exponential families and what they offer, especially wrt. Variational Inference

2.2.5 Explain the key concepts of MCMC (e.g. detailed balance, normalization) and the advantages disadvantages of MCMC vs Variational Inference

2.3 Extensions of Variational Inference

2.3.1 Explain SVI and its motivation and problems

2.3.2 What is a natural gradient and the motivation for it?

2.3.3 Explain Black Box Variational Inference

2.3.4 What are the disadvantages of Black Box Var. Inference?

2.3.5 Show variance reduction through control variates. What is one typical and simple choice for a control variate?

2.4 Dimensionality Reduction

2.4.1 Derive Factor Models, PCA, Probabilistic PCA

2.4.2 Derive EM for Factor Models

2.4.3 Show the relation between variance and reconstruction error

2.4.4 Show the relation between SVD and PCA

2.4.5 Explain the difference/similarities between PCA and linear least squares

2.4.6 Explain and derive the Eigenface Algorithm, name problems and results wrt. implementation and performance

2.4.7 Explain the Kernel Trick and derive Kernel PCA

2.4.8 What would be advantages and disadvantages wrt. the Eigenface Algorithm when using the Kernel trick?

2.5 Sequential Data

Our data aren't iid anymore. We account for that by conditioning on previous state. Simplest are first-order Markov models. We condition the current state on the previous one.

2.5.1 Explain HMM and Kalman Filter, especially Inference and Learning

2.5.2 Explain the difference between the Kalman Filter and the Rauch Tung Striebel smoother.

I didn't find this in the slides. Perhaps it's in Bishop. According to `https://en.wikipedia.org/wiki/Kalman_filter#Rauch%E2%80%93Tung%E2%80%93Striebel`, it's an efficient two-pass algorithm for fixed interval smoothing.

2.6 Score Functions

Score functions and the log derivative trick is explained here. `http://blog.shakirm.com/2015/11/machine-learning-trick-of-the-day-5-log-derivative-trick/`

2.6.1 Explain the concept and Score Function Estimators

See above.

2.6.2 Explain the log-derivative trick and how to derive: $\nabla_\theta \mathbb{E}_{x\sim p(x|\theta}\left[f(x)\right] = \mathbb{E}_x\left[f(x)\nabla_\theta \log p(x|\theta)\right]$

See above.

2.6.3 Relate the algorithm to policy gradients and discuss the problems of the algorithm

2.7 Variational Autoencoders

2.7.1 Explain the concept in detail and the motivation for distributional assumptions.

2.7.2 Explain the reparameterization trick.

2.7.3 Explain the main ideas and properties of SGD. How is it useful in the context of VAE and why do we need the reparameterization trick?

2.7.4 Show the derivation of VAE's from black box variational inference.

2.7.5 Explain the cost function of beta-VAE and outline 3 desired effects of the cost function compared to a classic VAE.

2.8 Deep Generative Models

2.8.1 Explain the cost function of GANs what are advantages and disadvantages of GANs compared to VAEs?

2.8.2 Explain the difficulty of evaluating generative models and especially their latent representation.

2.8.3 Explain ICA and how it relates to recent approaches for learning "disentangled" representations