Bayesian Hierarchical Classification

Dmitry Kondrashkin (HSE)
in collaboration with Dr. Christoph Lampert (IST Austria)

October 23, 2015

Hierarchical classification

Current approaches:

- Non-Bayesian: hierarchy of classifiers.
- Bayesian: hierarchy of priors.

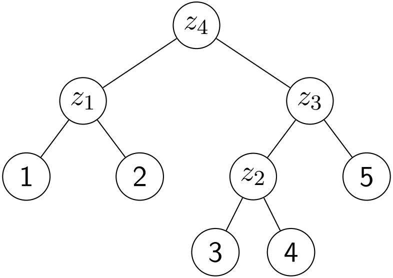
Our approach:

- Use information provided by hierarchy to adjust model complexity.

Notation

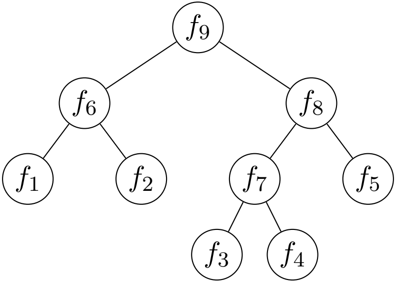
- $(X, Y) = \{(x_n, y_n)\}_{n=1}^{N}$,
- $y_n \in \{1, \ldots, K\}$,
- M is the number of nodes in the hierarchy tree,
- K is the number of classes (or leaf nodes).
- $Z = \{z_l\}_{l=1}^{M-K}$, $z_l \in \{0, 1\}$ is binary latent variable,
- $F = \{f_m\}_{m=1}^{M}$, f_m is Gaussian Process, i.e. $f_m \sim \mathcal{N}(0, K_m)$, $K_m \in \mathbb{R}^{N \times N}$,
- $\mathrm{path}(i)$ is the set of all the nodes in the path from i-th node to the root,
- $\mathrm{cl}(i)$ is the set of all the leaf nodes (classes) which are in the subtree with root in the i-th node.

Example: binary latent variables



- $K = 5$,
- $M = 9$,
- $Z = \{z_1, z_2, z_3, z_4\}$.

Example: Gaussian Processes



- $K = 5$,
- $M = 9$,
- $Z = \{z_1, z_2, z_3, z_4\}$,
- $F = \{f_1, f_2, \ldots, f_9\}$.

Probabilistic model

Complete likelihood:

$$p(Y, Z, F \mid X, \Theta) = \prod_{n=1}^{N} p(y_n, z_n \mid F) \prod_{m=1}^{M} \mathcal{N}(f_m \mid 0, K(\theta_m)).$$

- $\Theta = \{\theta\}_{m=1}^{M}$ are hyperparameters,
- $K(\cdot)$ is the kernel function.

Probabilistic model

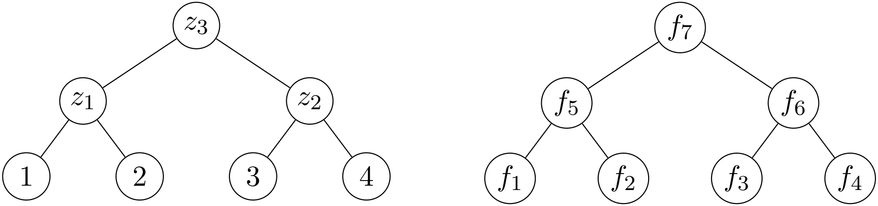
Likelihood for one object:

$$p(y_n = k, z_n \mid F) \propto \exp\left\{ f_{nk} \prod_{i \in \mathrm{path}(k)} (1 - z_i) \right.$$
$$\left. + \sum_{j \in \mathrm{path}(k)} (f_{nj} + \rho_j) z_j \prod_{i \in \mathrm{path}(j)} (1 - z_i) \right\}$$

- $\rho_k = -\ln |\,\mathrm{cl}(k)|$ is a penalty term,
- $z_i = 1$ means that all the classes from $\mathrm{cl}(i)$ are merged into one class.

Example

Consider the following class hierarchy:

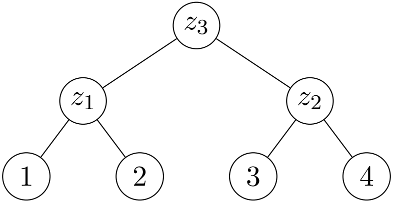


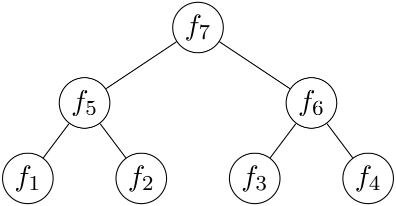
Probability for each class:

- $p(y=1) \propto \exp\{f_1(1-z_1)(1-z_3)+(f_5-\ln 2)z_1(1-z_3)+(f_7-\ln 4)z_3\}$,
- $p(y=2) \propto \exp\{f_2(1-z_1)(1-z_3)+(f_5-\ln 2)z_1(1-z_3)+(f_7-\ln 4)z_3\}$,
- $p(y=3) \propto \exp\{f_3(1-z_2)(1-z_3)+(f_6-\ln 2)z_2(1-z_3)+(f_7-\ln 4)z_3\}$,
- $p(y=4) \propto \exp\{f_4(1-z_2)(1-z_3)+(f_6-\ln 2)z_2(1-z_3)+(f_7-\ln 4)z_3\}$.

Example

Consider the following class hierarchy:



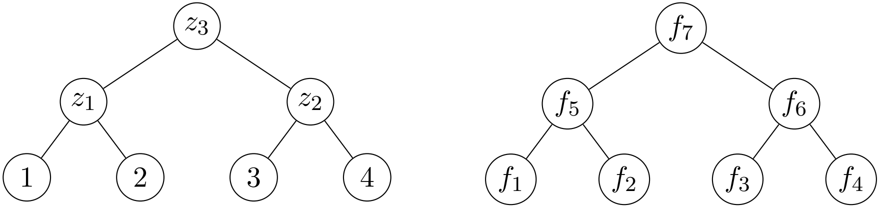


If all $z_i = 0$:

- $p(y = 1) \propto \exp\{f_1\}$,
- $p(y = 2) \propto \exp\{f_2\}$,
- $p(y = 3) \propto \exp\{f_3\}$,
- $p(y = 4) \propto \exp\{f_4\}$.

Example

Consider the following class hierarchy:

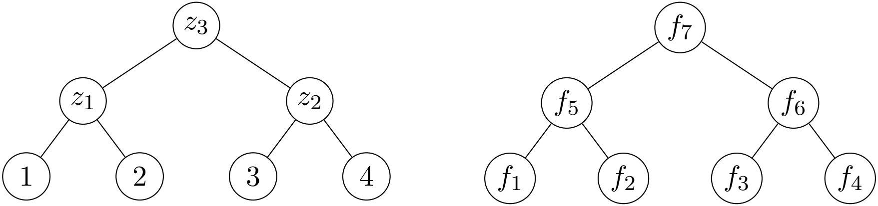


Probability for each class:

- $p(y=1) \propto \exp\{f_1(1-z_1)(1-z_3) + (f_5 - \ln 2)z_1(1-z_3) + (f_7 - \ln 4)z_3\}$,
- $p(y=2) \propto \exp\{f_2(1-z_1)(1-z_3) + (f_5 - \ln 2)z_1(1-z_3) + (f_7 - \ln 4)z_3\}$,
- $p(y=3) \propto \exp\{f_3(1-z_2)(1-z_3) + (f_6 - \ln 2)z_2(1-z_3) + (f_7 - \ln 4)z_3\}$,
- $p(y=4) \propto \exp\{f_4(1-z_2)(1-z_3) + (f_6 - \ln 2)z_2(1-z_3) + (f_7 - \ln 4)z_3\}$.

Example

Consider the following class hierarchy:

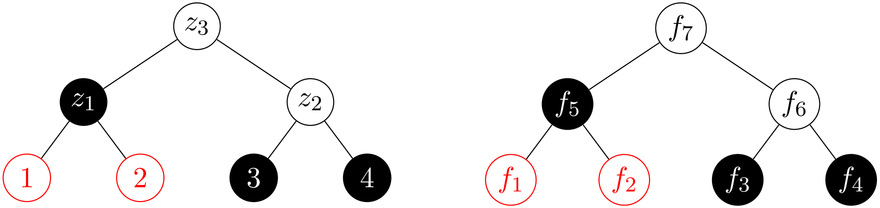


If all $z_i = 1$:

- $p(y = 1) \propto \exp\{f_7 - \ln 4\}$=0.25,
- $p(y = 2) \propto \exp\{f_7 - \ln 4\}$=0.25,
- $p(y = 3) \propto \exp\{f_7 - \ln 4\}$=0.25,
- $p(y = 4) \propto \exp\{f_7 - \ln 4\}$=0.25.

Example

Consider a configuration: $z_1 = 1, z_2 = 0, z_3 = 0$:



- $p(y = 1) \propto \exp\{f_5 + \ln 0.5\} = 0.5 \exp\{f_5\}$,
- $p(y = 2) \propto \exp\{f_5 + \ln 0.5\} = 0.5 \exp\{f_5\}$,
- $p(y = 3) \propto \exp\{f_3\}$,
- $p(y = 4) \propto \exp\{f_4\}$.

Inference: Variational Bayes

We use variational Bayes to approximate the posterior over Z and F:

$$\ln p(Y \mid X, \Theta) \geq \sum_Z \int q(Z)q(F) \ln \frac{p(Y, Z \mid F)p(F \mid \Theta)}{q(Z)q(F)} dF$$

- Here we approximate the posterior $p(F, Z) \approx q(F)q(Z)$,
- $p(F \mid \Theta)$ is the prior,

$p(Y, Z \mid F)$ is a softmax, thus the integral is intractable, there are two options to overcome this issue:

- Use local variational bounds like Jaakkola-Jordan bound to obtain a quadratic lower bound for the softmax to make this integral tracktable.
- Use stochastic optimization and reparametrization trick to compute gradient of the lower bound.

Inference: Variational Bayes

We would like to find $q(F) = \prod_{m=1}^{M} \mathcal{N}(f_m \mid \mu_m, \Sigma_m)$ and $q(Z)$ in the family of discrete distributions.

Variational Bayes update rules:

- $q(Z) \propto \exp\left\{\mathbb{E}_{q(F)} \ln p(Y, Z \mid F) p(F \mid \Theta)\right\}$
- $q(F) \propto \exp\left\{\mathbb{E}_{q(Z)} \ln p(Y, Z \mid F) p(F \mid \Theta)\right\}$

The issues:

- The first integration is intractable because of softmax in likelihood.
- We could compute $\mathbb{E}_{q(Z)}$ by explicit summation only if $|Z|$ is small.

Solving the first issue: local variational bound

1. bound log-sum-exp:

$$\ln \sum_{k=1}^{K} e^{x_k} \leq \alpha + \sum_{k=1}^{K} \ln(1 + e^{x-\alpha}), \forall \alpha \in \mathbb{R}$$

2. bound log-sigmoid using Jaakkola-Jordan bound:

$$\ln(1 + e^{x}) \leq \lambda(\xi)(x^2 - \xi^2) + (x - \xi)/2 + \ln(1 + e^{\xi}), \forall \xi \in \mathbb{R},$$

where $\lambda(\xi) = \frac{1}{2\xi}(1/(1 + e^{-\xi}) - 0.5)$.

Solving the second issue

Update rule for $q(F)$ after applying the bound:

$$\ln q(F) = \text{const} + \text{LogPrior} + \sum_{n=1}^{N}\sum_{k=1}^{K}\Bigg([y_n = k](f_{nk}\mathbb{E}_Z\tilde{z}_k^0 + \sum_{j\in\text{path}(k)}\tilde{f}_{nj}\mathbb{E}_Z\tilde{z}_k) - 0.5(f_{nk}\mathbb{E}_Z\tilde{z}_k^0 + \sum_{j\in\text{path}(k)}\tilde{f}_{nj}\mathbb{E}_Z\tilde{z}_k) - \lambda(\xi_{nk})(f_{nk}^2\mathbb{E}_Z\tilde{z}_k^0 + \sum_{j\in\text{path}(k)}\tilde{f}_{nj}^2\mathbb{E}_Z\tilde{z}_k + \ldots)\Bigg)$$

where

- $\tilde{f}_{nj} = f_{nj} + \rho_j$
- $\tilde{z}_k^0 = \prod_{j\in\text{path}(k)}(1 - z_j)$
- $\tilde{z}_k = z_k \prod_{j\in\text{path}(k)}(1 - z_j)$

Solving the second issue: message passing

Note that

$$\mathbb{E}_Z \tilde{z}_k^0 = \mathbb{E}_Z \prod_{j \in \mathrm{path}(k)} (1 - z_j) = q(z_{j_1} = 0, \ldots, z_{j_s} = 0)$$

$$\mathbb{E}_Z \tilde{z}_k = \mathbb{E}_Z z_k \prod_{j \in \mathrm{path}(k)} (1 - z_j) = q(z_k = 1, z_{j_1} = 0, \ldots, z_{j_s} = 0)$$

where $\{j_1, \ldots, j_s\} = \mathrm{path}(k)$. These are the marginals of $q(Z)$, they could be efficiently computed using message passing.

Inference: local bounds

Pros:

- We could derive closed form updates for $\xi_{nk}, \alpha_n, q(F)$ and $q(Z)$.
- We could use effective dynamic programming approach to deal with discrete distribution $q(Z)$.

Cons:

- Comlicated implementation.
- We introduce $N \times K + N$ additional variational parameters $\{\xi_{nk}\}_{n=1,k=1}^{N,K}$ and $\{\alpha_n\}_{n=1}^{N}$.
- The bound might be untight.

Inference: stochastic optimization

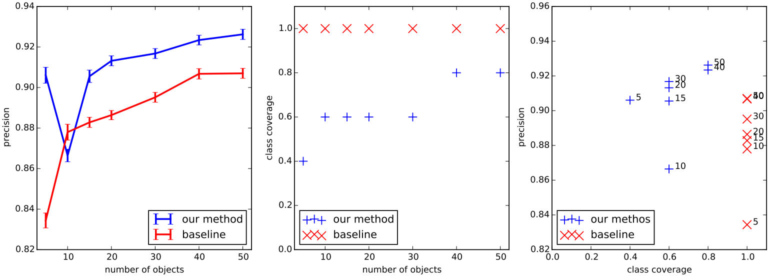
Pros:

- Very simple implementation using Theano.
- Using enough samples we could obtain very good approximation.

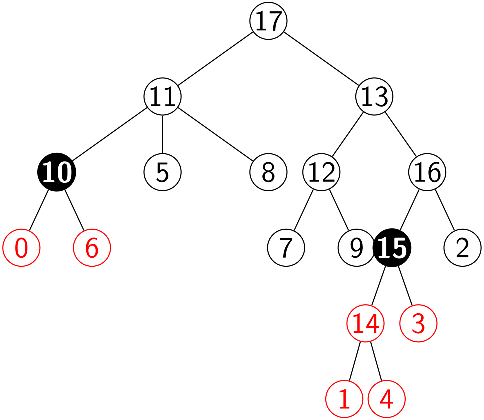
Cons:

- Reparametrization trick doesn't work for discrete distributions: only explicit summation available.
- Convergence issues.
- Slow.

Experiments: 10 ImageNet classes

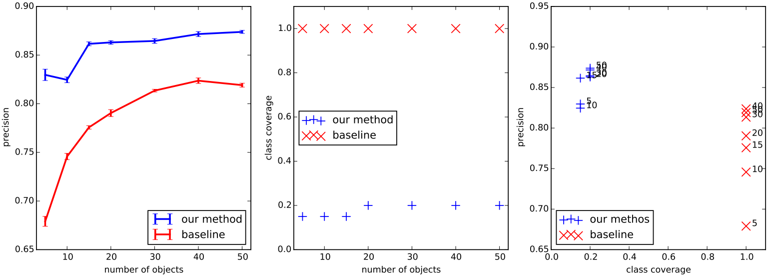


Experiments: 10 ImageNet classes, 5 objects per class



- 0 kit fox
- 1 English setter
- 2 Siberian husky
- 3 Australian terrier
- 4 English springer
- 5 grey whale
- 6 red panda
- 7 Egyptian cat
- 8 ibex
- 9 Persian cat

Experiments: 20 ImageNet classes



Conclusion

In this project we have used the following ideas:

- Marginal likelihood optimization to adjust model complexity.
- Local variational bounds to perform inference.
- Message passing to handle discrete distribution $q(Z)$.