Machine Learning 1
Lecture 13.1 - Combining Models
Bayesian Averaging vs Model Combination

Erik Bekkers

(Bishop 14.0, 14.1)
Regression with GP’s

› Combining models: (Bishop 4.1-4.4)

› **Bayesian model averaging vs. model combination methods**

› Committees:
  › Bootstrap aggregation
  › Random subspace methods
  › Boosting

› Decision trees

› Random forests
Regression with GP’s

› So far we have considered many different models for classification and regression

› It is often the case that overall performance can be improved by combining multiple models together in some way.

› Regression example: train $L$ different models and make predictions using the average of the predictions made by each model.

› Methods that are combined like this are called **committees**
Model combination

- Committee example: boosting

- Train multiple models in sequence

- Error function used to train a particular model depends on performance of previous models

- Boosting algorithms can lead to substantial improvements over individual models!
Alternative: model selection

- For each prediction, select one model to make a prediction.

- The choice of the model that is selected is a function of the input variables.

- Example 1: Decision trees! Selection process is a sequence of binary selections corresponding to the traversal of a tree structure.

- Example 2: Mixtures of experts. Soft selection of models for predictions.

\[ p(t | x) = \sum_{k=1}^{K} \pi_k(x)p(t | x, k) \]

One of the experts' mixing coefficients, which depends on \( x \).
Bayesian model averaging

VS

Model combination methods
Bayesian model averaging vs. combining models

- Let’s make sure we understand the difference between Bayesian model averaging and model combination methods

- We have already seen a model combination method for density estimation: Gaussian mixture models!

- Several Gaussians are combined probabilistically to produce the density $p(x)$

$$p(x) = \sum_{z} p(x \mid z) p(z)$$

- A binary latent variable $z$ that indicates which component of the mixture is responsible for generating the datapoint $x$.

- The model specifies

$$p(x, z) = p(x \mid z)p(z)$$

with

$$p(x \mid z_k = 1) = \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

$z \in \mathbb{R}^k, \ z \sim (0, 0, 1, \ldots )^T$ (one-hot enc.)
Combining models: GMM

- A Gaussian Mixture model specifies \( p(x, z) = p(x | z)p(z) \)
  \[
p(x | z_k = 1) = \mathcal{N}(x | \mu_k, \Sigma_k)
  \]

- Then the density over observed \( x \) is obtained by
  \[
p(x) = \sum_z p(x | z)p(z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x | \mu_k, \Sigma_k)
  \]

- For i.i.d data \( X = [x_1, \ldots, x_N]^T \) we have
  \[
p(X) = \prod_{n=1}^{N} p(x_n) = \prod_{n=1}^{N} \sum_{z_n} p(x_n, z_n)
  \]

- So each observed datapoint \( x_n \) has its own latent variable \( z_n \)!
Bayesian model averaging

- Suppose we have different models, indexed by $h = 1, \ldots, H$

- We also have prior probabilities $p(h)$

- Marginal distribution over the dataset is

  $$p(X) = \sum_{h=1}^{H} p(X | h)p(h)$$

- Interpretation: one model is responsible for generating the entire dataset

- $p(h)$ simply reflects our uncertainty which model is the correct model

- If dataset size increases, uncertainty is reduced: $p(h | X)$ becomes increasingly focused on one model
Contrast with model combination methods

- In Bayesian model averaging: the entire dataset is generated by a single model, we are just unsure which one it was!

- When we combine multiple models different data points can potentially be generated by different components/models!

- Example in GMM:
  - Take two datapoints $\mathbf{x}$ and $\mathbf{x}'$
  - They can be generated from different $\mathbf{z}$ and $\mathbf{z}'$
  - So they come from different model components!