Guided Tour of Machine Learning in Finance

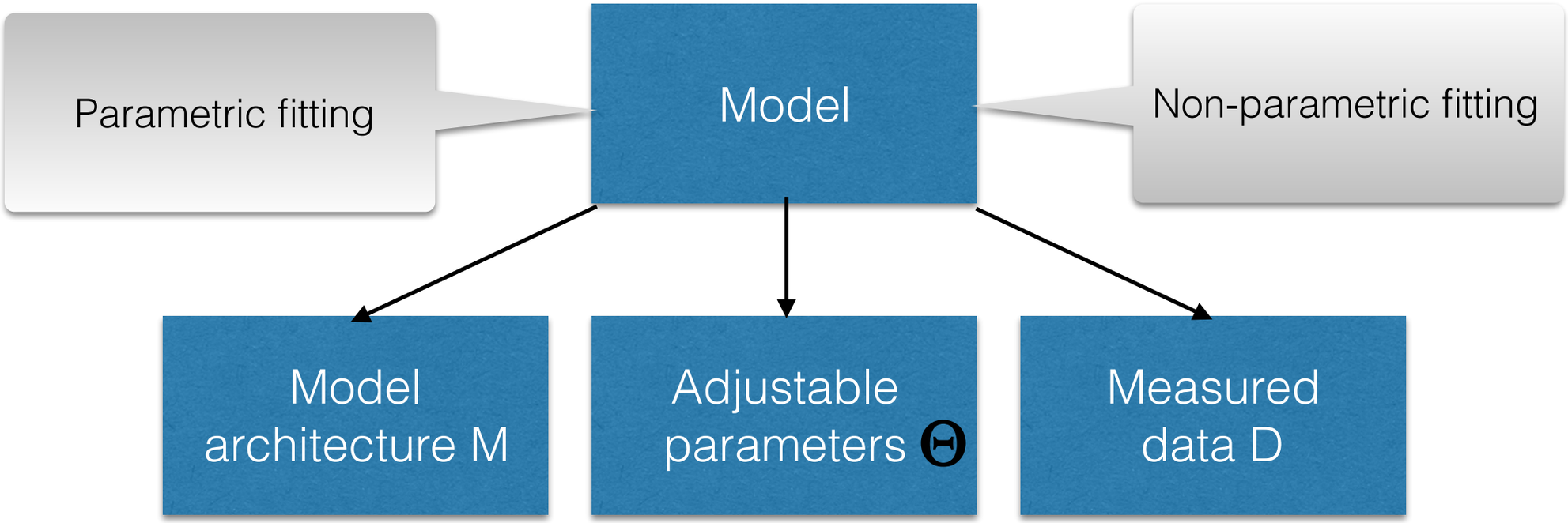
Week 2-Lesson 3-part 1:
Machine Learning with Probabilistic Models

Igor Halperin

NYU Tandon School of Engineering, 2017

ML as function fitting

All ML algorithms are about **fitting some (regularizer) loss function** $f(\mathbf{X},\Theta)$ to some data $\mathbf{D}$, where $\mathbf{X}$ is a vector of features, Θ is a vector of model parameters, and function $f(\mathbf{X}) = f(\mathbf{X},\Theta)$ belongs in some parametric family $\mathcal{F}_\Theta$. ML essentially amounts to a **model estimation** problem:



Parametric fitting: functions of particular form, with a small number of parameters

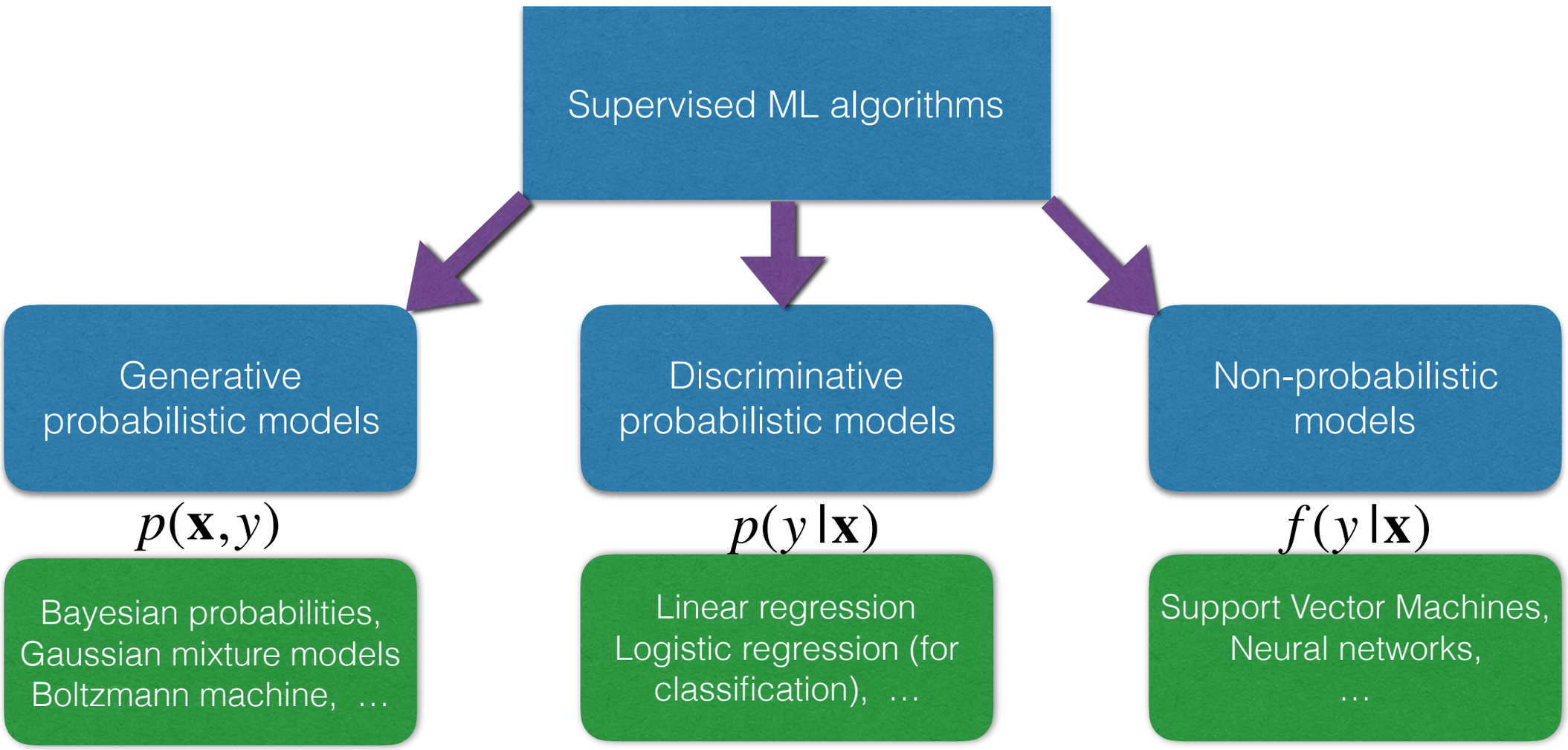
Non-parametric fitting: a very flexible function with many free parameters

Supervised Learning algorithms

Most, but not all, supervised ML algorithms amount to **estimating a probability distribution** $p(y \mid \mathbf{x})$

Example: Linear Regression is equivalent to a **discriminative probabilistic model**

$$p(y \mid \mathbf{x}) = \mathcal{N}(y; \Theta^T \mathbf{x}; \sigma^2)$$



ML training for probabilistic models

ML essentially amounts to a **model estimation** problem: What are the most probable values of parameters $\boldsymbol{\Theta}$, given the model $\mathbf{M} = f(\mathbf{X}, \boldsymbol{\Theta})$ and data $\mathbf{D}$? That is, we need to find $\boldsymbol{\Theta}$ that maximizes $p(\boldsymbol{\Theta} \mid \mathbf{D}, \mathbf{M})$
Use **Bayes' rule**: (assuming our model is a **probabilistic model!**)

$$\max_{\boldsymbol{\Theta}} p(\boldsymbol{\Theta} \mid \mathbf{D}, \mathbf{M}) = \max_{\boldsymbol{\Theta}} \frac{p(\mathbf{D} \mid \mathbf{M}, \boldsymbol{\Theta})\, p(\boldsymbol{\Theta} \mid \mathbf{M})}{p(\mathbf{D} \mid \mathbf{M})}$$

ML training for probabilistic models

ML essentially amounts to a **model estimation** problem: What are the most probable values of parameters $\boldsymbol{\Theta}$, given the model $\mathbf{M} = f(\mathbf{X}, \boldsymbol{\Theta})$ and data $\mathbf{D}$? That is, we need to find $\boldsymbol{\Theta}$ that maximizes $p(\boldsymbol{\Theta} \mid \mathbf{D}, \mathbf{M})$
Use **Bayes' rule**: (assuming our model is a **probabilistic model!**)

$$\max_{\boldsymbol{\Theta}} p(\boldsymbol{\Theta} \mid \mathbf{D}, \mathbf{M}) = \max_{\boldsymbol{\Theta}} \frac{p(\mathbf{D} \mid \mathbf{M}, \boldsymbol{\Theta})\, p(\boldsymbol{\Theta} \mid \mathbf{M})}{p(\mathbf{D} \mid \mathbf{M})}$$

$$= \max_{\boldsymbol{\Theta}} \frac{p(\mathbf{D} \mid \mathbf{M}, \boldsymbol{\Theta})\, p(\boldsymbol{\Theta} \mid \mathbf{M})}{\int p(\mathbf{D} \mid \mathbf{M}, \boldsymbol{\Theta})\, p(\boldsymbol{\Theta} \mid \mathbf{M})\, d\boldsymbol{\Theta}}$$

ML training for probabilistic models

ML essentially amounts to a **model estimation** problem: What are the most probable values of parameters $\boldsymbol{\Theta}$, given the model $\mathbf{M} = f(\mathbf{X}, \boldsymbol{\Theta})$ and data $\mathbf{D}$? That is, we need to find $\boldsymbol{\Theta}$ that maximizes $p(\boldsymbol{\Theta} \mid \mathbf{D}, \mathbf{M})$
Use **Bayes' rule**: (assuming our model is a **probabilistic model!**)

$$\max_{\boldsymbol{\Theta}} p(\boldsymbol{\Theta} \mid \mathbf{D}, \mathbf{M}) = \max_{\boldsymbol{\Theta}} \frac{p(\mathbf{D} \mid \mathbf{M}, \boldsymbol{\Theta})\, p(\boldsymbol{\Theta} \mid \mathbf{M})}{p(\mathbf{D} \mid \mathbf{M})}$$

$$= \max_{\boldsymbol{\Theta}} \frac{p(\mathbf{D} \mid \mathbf{M}, \boldsymbol{\Theta})\, p(\boldsymbol{\Theta} \mid \mathbf{M})}{\int p(\mathbf{D} \mid \mathbf{M}, \boldsymbol{\Theta})\, p(\boldsymbol{\Theta} \mid \mathbf{M})\, d\boldsymbol{\Theta}}$$

$$= \max_{\boldsymbol{\Theta}} \frac{Likelihood \times Prior}{Evidence}$$

ML training for probabilistic models

ML essentially amounts to a **model estimation** problem: What are the most probable values of parameters $\boldsymbol{\Theta}$, given the model $\mathbf{M} = f(\mathbf{X}, \boldsymbol{\Theta})$ and data $\mathbf{D}$? That is, we need to find $\boldsymbol{\Theta}$ that maximizes $p(\boldsymbol{\Theta} \mid \mathbf{D}, \mathbf{M})$
Use **Bayes' rule**: (assuming our model is a **probabilistic model!**)

$$\max_{\boldsymbol{\Theta}} p(\boldsymbol{\Theta} \mid \mathbf{D}, \mathbf{M}) = \max_{\boldsymbol{\Theta}} \frac{p(\mathbf{D} \mid \mathbf{M}, \boldsymbol{\Theta})\, p(\boldsymbol{\Theta} \mid \mathbf{M})}{p(\mathbf{D} \mid \mathbf{M})}$$

$$= \max_{\boldsymbol{\Theta}} \frac{p(\mathbf{D} \mid \mathbf{M}, \boldsymbol{\Theta})\, p(\boldsymbol{\Theta} \mid \mathbf{M})}{\int p(\mathbf{D} \mid \mathbf{M}, \boldsymbol{\Theta})\, p(\boldsymbol{\Theta} \mid \mathbf{M}) d\boldsymbol{\Theta}}$$

$$= \max_{\boldsymbol{\Theta}} \frac{Likelihood \times Prior}{Evidence}$$

Measures the match between the data and prediction of the model

Measures how well the model can describe the data

Prior beliefs about which values of the model are most reasonable

Control question

Q: Which statement below is correct:

A1. As Bayes' formula uses prior probabilities, Bayesian probabilities are too subjective to be taken seriously in any business that deals with money.
A2. The Evidence in Bayesian probability is the denominator in the Bayes' rule. As it does not depend on Theta, it should not matter if all we want to do is to find the best value of Theta.
A3. The Evidence is NEVER important for Bayesian statistics, as we can always work with un-normalized probabilities.

Correct answer: A2.

Maximum Likelihood Estimation (MLE)

Bayes' rule produced a general relation:

$$\max_{\Theta} p(\Theta \mid \mathbf{D}, \mathbf{M}) = \max_{\Theta} \frac{p(\mathbf{D} \mid \mathbf{M}, \Theta)\, p(\Theta \mid \mathbf{M})}{p(\mathbf{D} \mid \mathbf{M})} = \max_{\Theta} \frac{Likelihood \times Prior}{Evidence}$$

Special cases:

- The **model is fixed**, can drop conditioning on $\mathbf{M}$: $p(\mathbf{D} \mid \mathbf{M}) \rightarrow p(\mathbf{D})$

- The model is fixed, and a flat prior $p(\Theta \mid \mathbf{M}) = const$ is used:
  Maximum Likelihood Estimation (MLE)

$$\max_{\Theta} p(\Theta \mid \mathbf{D}, \mathbf{M}) = \max_{\Theta} p(\mathbf{D} \mid \mathbf{M}, \Theta)$$

Why MLE?

Maximum Likelihood Estimation (MLE):

$$\max_{\Theta} p(\Theta \mid \mathbf{D}, \mathbf{M}) = \max_{\Theta} p(\mathbf{D} \mid \mathbf{M}, \Theta)$$

Properties:

- **Consistent** estimator (i.e. $\hat{\Theta}_{N \to \infty} \to \Theta$) **if**:
  - The true distribution $p_{\text{data}}(\mathbf{x})$ lies within the family $\mathcal{F}_{\Theta}$
  - There is a unique value of Θ that corresponds to $p_{\text{data}}(\mathbf{x})$
- Asymptotically (as $N \to \infty$), has the **lowest possible MSE** among all consistent estimators (Cramer-Rao lower bound, 1945-1946)
- For **finite** N, biased estimators (e.g. a regularized MLE) can be more efficient (i.e. reach the same level of generalization error with a smaller number of samples N)

Maximum A-Posteriori (MAP) Estimation

Bayes' rule produced a general relation:

$$\max_{\Theta} p(\Theta \mid \mathbf{D}, \mathbf{M}) = \max_{\Theta} \frac{p(\mathbf{D} \mid \mathbf{M}, \Theta)\, p(\Theta \mid \mathbf{M})}{p(\mathbf{D} \mid \mathbf{M})} = \max_{\Theta} \frac{Likelihood \times Prior}{Evidence}$$

Special cases:

- The **model is fixed**, can drop conditioning on $\mathbf{M}$: $p(\mathbf{D} \mid \mathbf{M}) \rightarrow p(\mathbf{D})$

- The model is fixed, and a flat prior $p(\Theta \mid \mathbf{M}) = const$ is used:
  Maximum Likelihood Estimation (MLE)

$$\max_{\Theta} p(\Theta \mid \mathbf{D}, \mathbf{M}) = \max_{\Theta} p(\mathbf{D} \mid \mathbf{M}, \Theta)$$

- The model is fixed, and a non-flat prior $p(\Theta \mid \mathbf{M})$ is used:
  Maximum A-Posteriori (MAP) estimation

$$\max_{\Theta} p(\Theta \mid \mathbf{D}, \mathbf{M}) = \max_{\Theta} p(\mathbf{D} \mid \mathbf{M}, \Theta)\, p(\Theta \mid \mathbf{M})$$

MLE and least squares loss

Assume we want to fit *N* noisy measurements of quantity y_n as a function of variable $\mathbf{x_n}$: $y_n = f(\mathbf{x_n}, \Theta) + \varepsilon_n$

Assuming that errors ε_n are i.i.d. and have a Gaussian distribution with variance σ_n^2, the likelihood is

$$p(\mathbf{D} \mid \mathbf{M}, \Theta) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{\left(y_n - f(\mathbf{x_n}, \Theta)\right)^2}{2\sigma_n^2}}$$

MLE and least squares loss

Assume we want to fit *N* noisy measurements of quantity y_n as a function of variable $\mathbf{x_n}$: $y_n = f(\mathbf{x_n}, \Theta) + \varepsilon_n$

Assuming that errors ε_n are i.i.d. and have a Gaussian distribution with variance σ_n^2, the likelihood is

$$p(\mathbf{D} \mid \mathbf{M}, \Theta) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{\left(y_n - f(\mathbf{x_n}, \Theta)\right)^2}{2\sigma_n^2}}$$

Maximization of $p(\mathbf{D} \mid \mathbf{M}, \Theta)$ is equivalent to minimization of **negative log-likelihood** (**NLL**):

$$\min_{\Theta} \left(-\log\, p(\mathbf{D} \mid \mathbf{M}, \Theta)\right) = \min_{\Theta} \sum_{n-1}^{N} \frac{\left(y_n - f(\mathbf{x_n}, \Theta)\right)^2}{2\sigma_n^2} + \frac{1}{2}\log(2\pi\sigma_n^2)$$

MLE and least squares loss

Assume we want to fit *N* noisy measurements of quantity y_n as a function of variable $\mathbf{x_n}$: $y_n = f(\mathbf{x_n}, \Theta) + \varepsilon_n$

Assuming that errors ε_n are i.i.d. and have a Gaussian distribution with variance σ_n^2, the likelihood is

$$p(\mathbf{D} \mid \mathbf{M}, \Theta) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{\left(y_n - f(\mathbf{x_n}, \Theta)\right)^2}{2\sigma_n^2}}$$

Maximization of $p(\mathbf{D} \mid \mathbf{M}, \Theta)$ is equivalent to minimization of **negative log-likelihood** (**NLL**):

$$\min_{\Theta} \left(-\log\, p(\mathbf{D} \mid \mathbf{M}, \Theta)\right) = \min_{\Theta} \sum_{n-1}^{N} \frac{\left(y_n - f(\mathbf{x_n}, \Theta)\right)^2}{2\sigma_n^2} + \frac{1}{2}\log(2\pi\sigma_n^2)$$

When variances σ_n^2 are constant, this is equivalent to a **minimum least squares error (MSE)** function

$$\min_{\Theta} \sum_{n-1}^{N} \left(y_n - f(\mathbf{x_n}, \Theta)\right)^2$$

Becomes linear regression if

$$f(\mathbf{x_n}, \Theta) = \sum_k \Theta_k x_{nk} = \Theta^T \mathbf{x}$$

Kullback-Leibler (KL) divergence

Another interpretation of MLE : minimization of the KL divergence between the model distribution $p_{\text{model}}(\mathbf{x},\Theta)$ and the true distribution $p_{\text{data}}(\mathbf{x})$:

$$D_{KL}\left(\mathrm{p}_{\text{data}} \,\|\, p_{\text{model}}\right) = \mathbb{E}_{x\sim p_{data}}\left[\log \frac{p_{\text{data}}(\mathbf{x})}{p_{\text{model}}(\mathbf{x})}\right]$$

$$= \mathbb{E}_{x\sim p_{data}}\left[\log p_{\text{data}}(\mathbf{x}) - \log p_{\text{model}}(\mathbf{x})\right]$$

$$= \underbrace{\mathbb{E}_{x\sim p_{\text{data}}}\left[\log p_{\text{data}}(\mathbf{x})\right]}_{\textit{independent of } p_{\text{model}}} - \mathbb{E}_{x\sim p_{\text{data}}}\left[\log p_{\text{model}}(\mathbf{x})\right]$$

$$\simeq -\mathbb{E}_{x\sim \hat{p}_{\text{data}}}\left[\log p_{\text{model}}(\mathbf{x})\right] + \ldots = -\sum_{n} \log p_{\text{model}}(\mathbf{x}_n,\Theta)$$

KL divergence between two distributions $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ measures their dissimilarity:

$$D_{KL}\left(p_1 \,\|\, p_2\right) \geq 0$$

$$D_{KL}\left(p_1 \,\|\, p_2\right) = 0 \quad \textit{iff } p_1(\mathbf{x}) = p_2(\mathbf{x})$$

KL divergence is widely used in ML!

MAP and regularization

Assume we want to fit *N* noisy measurements of quantity y_n as a function of variable $\mathbf{X}_n$:

$$\max_{\Theta} \; p(\Theta \mid \mathbf{D}, \mathbf{M}) = \max_{\Theta} \; p(\mathbf{D} \mid \mathbf{M}, \Theta)\, p(\Theta \mid \mathbf{M})$$

Negative log-likelihood for our problem $y_n = f(\mathbf{X}_n, \Theta) + \varepsilon_n$:

$$\min_{\Theta} \left(-\log \; p(\mathbf{D} \mid \mathbf{M}, \Theta) \right) = \min_{\Theta} \sum_{n-1}^{N} \frac{\left(y_n - f(\mathbf{x_n}, \Theta) \right)^2}{2\sigma_n^2} - \log p(\Theta \mid \mathbf{M}) + \ldots$$

Training error function

Regularizer

MAP and regularization

Recall the definition of the Maximum A-Posteriori (MAP) Estimation of model parameters:

$$\max_{\Theta} \; p(\Theta \mid \mathbf{D}, \mathbf{M}) = \max_{\Theta} \; p(\mathbf{D} \mid \mathbf{M}, \Theta)\, p(\Theta \mid \mathbf{M})$$

Negative log-likelihood for our problem $y_n = f(\mathbf{x_n}, \Theta) + \varepsilon_n$:

$$\min_{\Theta} \left(-\log \; p(\mathbf{D} \mid \mathbf{M}, \Theta) \right) = \min_{\Theta} \sum_{n-1}^{N} \frac{\left(y_n - f(\mathbf{x_n}, \Theta) \right)^2}{2\sigma_n^2} - \log p(\Theta \mid \mathbf{M}) + \ldots$$

Training error function

Regularizer

Examples:

- Gaussian prior $\sim e^{-\lambda(\theta - \theta_0)^2} \quad \Rightarrow \quad L_2$ regularization $\lambda \left|\left| \Theta - \Theta_0 \right|\right|_2$
- Laplace prior $\sim e^{-\alpha|\theta - \theta_0|} \quad \Rightarrow \quad L_1$ regularization $\lambda \left|\left| \Theta - \Theta_0 \right|\right|_1$

Control question

Q: Select all correct statements:

1. Linear Regression with a MSE error is equivalent to a Linear Probabilistic model with a constant Gaussian noise.
2. Minimization of the negative log-likelihood function within the MLE method is equivalent to minimization of the KL-divergence between the data and model distributions.
3. The KL-divergence of two distributions $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ is equal to the difference of entropies of these two distributions.
4. The KL-divergence is a measure of dissimilarity between two distributions. It is always non-negative, and is equal zero iff $p_1(\mathbf{x}) = p_2(\mathbf{x})$
5. The relative entropy is obtained from the KL-divergence by subtracting the diverging part.

Correct answers: 1, 2, 4.