Advanced Machine Learning

Lectures

1 - Introduction

- Measurement space, features, typical learning problems, key concepts, what you should know
- Supervised vs unsupervised learning, generative vs discriminative modeling

2 - Representations

- Expected risk (R): conditional and total expected risk
- Empirical risk (R^): training error, empirical risk minimizer, test error
  - Distinguish between test error and expected risk
- Taxonomy of data, object space, measurement
  - Monadic, dyadic (e.g. pairwise), polyadic
- Scales
  - Nominal (categorical): qualitative, but without quantitative measurements
  - Ordinal: measurement values are meaningful only with respect to other measurements, i.e., the rank order of measurements carries the information, not the numerical differences
  - Quantitative scale
    - Interval: the relation of numerical differences carries the information. Invariance w.r.t. translation and scaling
    - Ratio: zero value of the scale carries information but not the measurement unit
    - Absolute: absolute values are meaningful
- Mathematical spaces: topological, metric, Euclidean vector, metrizable
- Probability spaces: elementary event, sample space, family of sets, algebra of events, probability of events, probability model (triplet)
  - **Stackexchange**: Where a distinction is made between probability function and density, the pmf applies only to discrete random variables, while the pdf applies to continuous random variables
  - **ml2016tutorial1**: Note: Expected value =/= Most likely value
  - Describing dependencies in data by covariance is equivalent to approximation of data distribution by a Gaussian model.

3 - Density Estimation in Regression: Parametric Models

- Modeling assumptions for regression, different approaches, Bayesianism and frequentism
- Maximum Likelihood Estimation, ML estimation for normal distributions
  - Procedure: Find the extremum of the log-likelihood function
  - **Wikipedia**: Under the additional assumption that the errors are normally distributed, ordinary least squares (OLS) is the maximum likelihood estimator.

    - **Wikipedia**: Gauss-Markov Theorem states that in a linear regression model in which the errors have expectation zero, are uncorrelated and have equal variances, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator, provided it exists. The errors do not need to be normal, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and homoscedastic with finite variance).
    - **ml2016tutorial3**: Note that if we don't know the real value of μ, we can use its obtained prediction μ^ to calculate σ^, however in this case σ^ would be biased, i.e. σ^ =/= σ_{true}.
    - The James—Stein estimator is a biased estimator of the mean of Gaussian random vectors. It can be shown that the James—Stein estimator dominates the "ordinary" least squares approach, i.e., it has lower mean squared error. It is the best-known example of Stein's phenomenon.
    - Maximum likelihood estimation of variance is biased, but it is nevertheless consistent.
- Rao-Cramer inequality, Fisher information, score etc.
    - **Wikipedia**: In its simplest form, the bound states that the variance of any unbiased estimator is at least as high as the inverse of the Fisher information.
    - **Wikipedia**: An unbiased estimator which achieves this lower bound is said to be (fully) efficient. Such a solution achieves the lowest possible mean squared error among all unbiased methods, and is therefore the minimum variance unbiased (MVU) estimator.
    - **Wikipedia**: The Cramér–Rao bound can also be used to bound the variance of biased estimators of given bias. In some cases, a biased approach can result in both a variance and a mean squared error that are below the unbiased Cramér–Rao lower bound
- Importance of the Maximum Likelihood Method, realizable model
- Summary of MLEs
    - Consistency, equivariance, asymptotic efficiency, asymptotic normality
- Bayesian Learning, on normal distribution, recursive Bayesian estimation
    - **Exercise 2**: Having determined the functional form of the prior and likelihood, we want to compute the posterior. Doing it analytically can be hard in general, but it is easy if the prior and likelihood form a conjugate pair. Then the posterior will have the same functional form as the prior, only the parameters differ.
    - **Wikipedia**: In Bayesian probability theory, if the posterior distributions are in the same probability distribution family as the prior probability distribution, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function
    - **ml2016tutorial3**: Conjugate priors:
        - the gamma distribution is conjugate to the exponential distribution
        - the normal distribution is conjugate to the normal one
- ML-Bayes estimation differences
    - The maximum likelihood method only estimates the parameters μ^, σ^, but not the distribution!
    - **ml2016tutorial3**: simple linear regression corresponds to MLE, regularized linear regression corresponds to MAP.
- Schematic behaviour of bias and variance

4 - Regression

- Linear regression models, least squares, residual sum of squares (RSS)

  - **ml2016tutorial2**: We can have several different models like linear, quadratic or logarithmic. Note that they are all linear in parameters β_j , can be solved by linear regression.
  - **ml2016tutorial2**: Note that if the number of features is greater than number of samples X^TX might become singular
  - **ml2016tutorial2**: Closed form solution: set first derivative to zero (since loss is a convex function), check if the second derivative is positive definite
  - **Wikipedia**: In regression analysis, the term mean squared error is sometimes used to refer to the unbiased estimate of error variance: the residual sum of squares divided by the number of degrees of freedom
- Optimality of the least squares estimate, Gauss-Markov Theorem
- Bias/variance tradeoff, bias/variance decomposition
  - Objective: Minimize bias and variance simultaneously - usually impossible
  - Small data sets and large *C*: variance large, bias small
  - Large data sets and small *C*: variance small, bias large
  - The optimal tradeoff between bias and variance is achieved when we avoid both underfitting (large bias) and overfitting (large variance)
  - Outlook: Ensemble methods seem to avoid the bias/variance tradeoff since they lower variance while keeping the bias fixed. Note: The Rao-Cramer inequality defines a lower bound for variance reduction by ensemble averaging (no free lunch)
  - Error decomposition
    - Noise: Irreducible error, lower bound on the expected error
    - Variance: Error due to variance in the training dataset
    - Bias: Error due to various reasons (model complexity, wrong model, etc.)
- Solutions to overfitting: regularization (choice of prior in Bayesian framework and MAP estimator), model selection based on generalization error estimate (e.g. cross-validation), ensembles of classifiers
- Regularization = Bayesian MAP estimates, ridge regression (Tikhonov regularization, weight decay) and LASSO, Singular Value Decomposition (SVD), Ridge vs LASSO, Shrinkage Methods
  - **ml2016tutorial2**: Even if X^TX is singular, $X^TX + \lambda I$ is invertible if λ is greater than zero (ridge regression solution adds a positive constant to the diagonal)
  - LASSO estimates are known to be sparse with few coefficients non-vanishing (the LSE error surface hits often the corners of the constraint surface)
  - **ml2016tutorial2**: No closed form solution with Lasso, but the problem is still convex. When C decreases, some coefficients will be exactly zero (this is not the case for Ridge regression).
  - Idea behind shrinkage: When white noise is added to the data then all Fourier coefficients are increased by a constant on average. Shrink all coefficients by the estimated noise amount to derive a robust predictor
  - **ml2016tutorial2**: Ridge pushes all the coefficients together towards zero, either all of them are zero or none of them is zero
- Nonlinear regression, basis expansion, splines

5 - Gaussian Processes of Regression

- Bayesian linear regression, model inversion, moments of Bayesian linear regression
  - **ml2016tutorial4**: To get a Bayesian point estimate, take the maximum a posteriori (MAP) estimate of the posterior distribution

    - **ml2016tutorial4**: Prior for Î¸ centered at 0: encourages Î¸ to be small
- Gaussian processes, kernelized linear regression
    - Kernel functions specify the degree of similarity between any two data points
    - The kernel functions encode our assumptions about the function which we wish to learn, e.g. its smoothness. Different kernel functions can be used to obtain very different models
    - **ml2016tutorial4**: If we can represent k(x,y) = φ(x)^{T}φ(y) for some function φ(.), then k(x,y) is a kernel function
- Kernel properties, Gram matrix, examples of kernel functions, kernel engineering, kernels beyond R^D, prediction by Gaussian processes, predictive density, model averaging
    - Unbiased estimators remain unbiased after averaging
    - Combining regressors: Variance reduction by a factor of 1/B - main effect if bias remains unchanged
    - **Wikipedia**: Intuitively, the polynomial kernel looks not only at the given features of input samples to determine their similarity, but also combinations of these. In the context of regression analysis, such combinations are known as interaction features
    - **Tufts**: Note that Mercer's theorem allows us to work with a kernel function without knowing which feature map it corresponds to or its relevance to the learning problem. This has often been used in practical applications.
    - **Intro. to ML**: Kernels provide a principled way of deriving nonparametric models from parametric ones
    - **Duvenaud**: Multiplying kernels can be thought as an AND operation and adding them as an OR operation.
    - RBF kernel (exponentiated quadratic)
        - **Duvenaud**: The lengthscale *l* determines the length of the 'wiggles' in your function. In general, you won't be able to extrapolate more than *l* units away from your data.
        - **Stackexchange**: Small lengthscale value means that function values can change quickly, large values characterize functions that change only slowly. Lengthscale also determines how far we can reliably extrapolate from the training data.
    - A stationary covariance function is a function of *x-x'* and it is invariant to the translations in the input space.
    - If the covariance function is only a function of |*x-x'*|, then it is called isotropic. It is thus invariant to all rigid motions, i.e. rotation, translation, reflection and their combinations.

6 - Numerical Estimation Techniques

- Bias variance tradeoff: *R = bias^2 + variance + noise*
    - High bias indicates underfitting -> increase model complexity
    - High variance indicates overfitting -> decrease model complexity
    - Ideal model should have low bias and low variance
- Numerical Estimation Techniques: cross-validation, bootstrap. jackknife, Neyman Pearson test, AIC, BIC...
- Core problem of statistical inference: true risk vs empirical risk, f-opt vs f-hat
- Cross-validation
    - **Intro. to ML**: Using one test set leads to overfitting on the test data.

    - **Intro. to ML**: Cross-validation only works if the data is i.i.d.
    - K-fold cross validation: There is a systematic tendency to underfit. Because we only use K-1 / K percent of the available data, the models are not as complex as they could be using the full data.
    - Leave-one-out cross validation (LOOCV): The leave-one-out method is unbiased w.r.t. the true prediction error but the variance can be very large due to highly correlated training sets (bias-variance dilemma).
    - **Tutorial 3**: LOOCV is deterministic, computationally expensive, has low bias and high variance
    - **Tutorial 3**: For linear models and quadratic error, there exist methods such that the models do not have to be retrained K times
    - Engineer's solution: choose K in K-fold cross validation as min{sqrt(n), 10}
    - Model fitting (training), model selection (validation), model assessment (test)
    - Model selection and model assessment both rely upon an estimation of the generalization error
    - **Tutorial 3**: The One-Standard Rule: Pick the simplest model f_i whose CV error is no more than 1 standard error above the lowest CV error over all models
    - Cross-validation “sacrifices” samples for validation which yields a bias towards too simple models.
    - Cross-validation is "okay" when you have much data.
- Bootstrapping
    - Idea: Estimate the uncertainty of a statistic S(Z) by resampling with replacement
    - Bootstrap sampling is consistent, but bootstrap estimates often show a too small bias (B = 200 is sufficient in most cases to estimate the error).
    - Bootstrap works if the deviation between empirical and bootstrap estimator (^F star being the Bootstrap CDF) converges in probability to the deviation between true parameter value and the empirical estimator
    - Naïve Bootstrap: Error estimation is too optimistic as there is overlap of training and test sets.
    - Leave-one-out bootstrap: Leave-one-out bootstrap solves the naïve bootstrap overfitting problem. However, as in CV, less data is used for training: on average, 0.632 x n samples. The introduced bias is similar as for 2-fold CV. (pessimistic estimate of the expected prediction error)
    - .632 bootstrap: This reduces the bias of leave-one-out bootstrap.
    - .632+ bootstrap: Also consider estimating the variability of the error rate estimate.
- Sketching distributions: moment methods (e.g. estimating statistics using CV or bootstrap), graphical sketch (e.g. box-plots, quartiles), density estimation (probability distribution)
- Jackknife
    - A smart way of de-biasing an estimator. (Especially if we mainly care about the bias, although we might introduce more variance)
    - Jackknife debiasing: Construction of an estimator with small bias (Bias corrected estimators can have a considerably larger variance than uncorrected estimators!)
- Neyman-Pearson Test, likelihood ratio test
    - Select a decision rule which minimizes the error of the second kind while fixing the error of the first kind!
    - **Wikipedia**: The power of a binary hypothesis test is the probability that the test rejects the null hypothesis (H_0) when a specific alternative hypothesis (H_1) is true. The statistical power ranges from 0 to 1, and as statistical power increases, the probability of making a type II error (wrongly failing to reject the null) decreases.

    - **Wikipedia**: The Neyman–Pearson lemma states that this likelihood ratio test is the most powerful among all level α tests for this problem.
- Complexity-Based Model Selection, Occam's razor, objective = loss + model complexity
- Bayesian Information Criterion (BIC)
    - **Wikipedia**: approximation is only valid for sample size n much larger than the number k of parameters in the model.
    - **Wikipedia**: the BIC cannot handle complex collections of models as in the variable selection (or feature selection) problem in high-dimension.
    - May impose too strong of a penalty on complex models (for small n)
- Minimum Description Length (MDL)
    - MDL and BIC are consistent as a model and asymptotically equivalent.
- Akaike Information Criterion (AIC)
    - Approximates the Kullback-Leibler (KL) divergence between the true model and the estimate.
    - In a sense, KL measures how much information do you lose when you use p^ instead of p.
    - AIC is asymptotically equivalent to Leave-one-out cross-validation for ordinary linear regression models (and mixed effect models). It can also be derived in a Bayesian framework, but with a different prior than for BIC (BIC has a uniform prior, which may not be very sensible)
    - The penalty on model complexity is smaller for AIC, as it misses the factor log(n). For small sample sizes, AIC has the tendency to select large models, i.e. to overfit, while BIC has the tendency to underfit
- Takeuchi Information Criterion (TIC)
    - Correction of AIC when true model is not an element of the model class. TIC reduces to AIC if the true model is an element of the model class.
    - For consistent models trace will be calculated on an identity matrix and be equal to the number of parameters k.
- Summary
    - With BIC, AIC, etc. we can train on the whole data (unlike CV and such) but we have to make some modeling assumptions.
    - Better to use BIC with more data and AIC with less data.

7 - Design of Linear Discriminant Functions

- Problem of Classification (Terminology, motivation)
    - Doubt and outlier classes, classification error (Indicator function, expected risk)
- Loss function
    - Weighted loss, asymmetric loss, 0-1 loss, its conditional risk, surrogate loss functions (exponential, logistic, hinge, epsilon-insensitive)
    - **cs229**: 0-1 loss is discontinuous, non-convex and NP-hard to minimize.
- Generative vs. Discriminative modeling
- Why Linear Classifiers?
    - Historical reasons, computational efficiency, statistical simplicity, optimality for Gaussian with equal covariances
    - **Wikipedia**: Finding a quadratic classifier for the original measurements would then become the same as finding a linear classifier based on the expanded measurement vector.
- Generalized Linear Discriminative Functions

    - Linear vs. Quadratic, Nonlinear feature transform, normalization, need for classifier margin
- Algorithms for learning the weight vector (What is the optimal learning rate)
    - Gradient descent, 2nd order algorithm with Taylor expansion
    - Newton's algorithm, Newton's rule
- Perceptron Algorithm
    - Perceptron Criterion (sum of violating projections), Perceptron Rule
    - Batch version, fixed-increment single sample
    - Perceptron convergence, bound on the number of steps
    - Limitations, credit assignment
    - Examples that are almost orthogonal (very near) to solution vector are difficult to learn.
    - **Intro. to ML**: Is just stochastic gradient descent (SGD) on the Perceptron loss function
- WINNOW Algorithm (exponential updates)
- Different cost functions
- Generative modeling approach
    - Class conditional density, density estimation, parameter estimation
    - Parametric and non-parametric statistics, statistical learning theory
- Bayes Rule for known densities and parameters
    - Bayes Optimal Classifier (corresponds to MAP), its derivation
    - **Auckland**: No other learner using the same hypothesis space and same prior knowledge can outperform this method on average.
- Outliers
- MAP examples (Class conditional densities and posteriors, errors in dichotomies)
- Gaussian Bayes Classifier
- Fisher's Linear Discriminant Analysis (classification by dimension reduction)
    - **ml2016tutorial3**: maximize the distance between classes while keeping distance within classes small (minimize class overlap)
- "Breakpoint of an estimator"
    - Percentage (number) of samples which, when you let them go to infinity, can make your estimator arbitrarily bad
    - Examples for Fisher's LDA, mean estimator, median estimator
- Separability measured by sample means
    - Scatter matrices, Fisher's Separation Criterion
    - Generalized Eigenvalue Problem, Unscaled solution
    - Multiple discriminant analysis, ambiguous regions

8 - Support Vector Machines

- **Tutorial 5**: Why maximize the margin? VC theory indicates that it is the right thing to do, i.e. good generalization properties.
- **Tutorial 5**: Soft-margin SVM motivation: Non-linearly-separable datasets, non-separable datasets, sensitivity to outliers
- **Tutorial 5**: Real-world SVM implementations usually combine three techniques:
    - Maximum margin classifier
    - Soft margin technique
    - Kernel trick

- Slack variables (ξ_i's) are sample specific relaxations on the margin, where they help with better classification.
- Linear penalty w.r.t. slack variables prefer a sparse solution. You either commit to a variable or you don't.
- Even when the problem is solvable, we might go for a little bit of slack to obtain more robust solutions.
- **Wikipedia**: In statistics, the Huber loss is a loss function used in robust regression, that is less sensitive to outliers in data than the squared error loss.

9 - Structural SVMs

- Ideas: linear classifier with margin, nonlinear transformation in kernel space
- Optimization with constraints, Lagrangian Optimization Theory

10 - Ensemble Methods for Classifier Design

- Boosting and bagging are empirical approaches to Bayesianist philosophy. Bayesian inference by model averaging: Keep all classifiers according to their weights (probabilities). (Compare with the ERM principle)
- Variance only goes down by averaging when the covariance matrix is sufficiently small.
- Weak Learners Used for Bagging or Boosting: Stumps, decision trees, multi-layer perceptrons, radial basis functions
- Diversity: Using different subsets of data, using different features, decorrelating classifiers during training
- **Wikipedia**: A decision stump is a machine learning model consisting of a one-level decision tree.
- **Wikipedia**: Bootstrap aggregating, often abbreviated as bagging, involves having each model in the ensemble vote with equal weight. In order to promote model variance, bagging trains each model in the ensemble using a randomly drawn subset of the training set.
- **Salakhutdinov**: In Bayesian model averaging just one model is responsible for generating the whole dataset, the distribution over h reflects our uncertainty as to which model that is. As we observe more data, the uncertainty decreases, and the posterior p(h|X) becomes focused on just one model.
- **Salakhutdinov**: Bagging almost always helps with regression, but even with unstable learners it can hurt in classification. If we bag a poor and unstable classifier we can make it horrible.
- First compare, then bag approach: calculate error differences of two sets of classifiers, pairwise on same bootstrap samples.
- Why does bagging work?: small covariances (different subsets for training), similar variances (estimators behave similarly), biases are weakly affected
- Random forest strategy for bagging trees: Grow a sufficiently deep decision tree to reduce bias -> noisy classifier -> average to enhance robustness, classify with majority vote.
    - Common splitting criteria are entropy, Gini index and misclassification rate.
    - Developer does not specify the exact classification rule, it comes from the diversity mechanism in the averaging of the algorithm.
    - Understandability is not necessary, what necessary is that we can control the solutions.
- AdaBoost algorithm

  - The minimizer of the exponential loss function (w.r.t. true distribution) is shown to be the log-odds of the class probabilities.
  - The discrete AdaBoost algorithm builds an additive logistic regression model via Newton-like updates for minimizing expected loss.
- **Wikipedia**: Specifically, if one considers AdaBoost as a generalized additive model and then applies the cost functional of logistic regression, one can derive the LogitBoost algorithm.
- **cs229**: Random forest is a tree-based technique that uses a high number of decision trees built out of randomly selected sets of features. Contrary to the simple decision tree, it is highly uninterpretable but its generally good performance makes it a popular algorithm.

11 - Probably Approximately Correct Learning

- Probability of large excess error problems:
  - We could be unlucky with the training dataset
  - The Bayes classifier is not in our hypothesis class
  - It is a random variable in training dataset Z and not a scalar quality measure
- Strategy should be to average over all sample sets Z. We have to look for a probabilistic statement, don't expect an exact result.
- Sample Complexity: smallest sample size that allows us to find a hypothesis with error at most ε with probability at least 1-δ for all distributions and all target concepts in C.
- Goal: derive error bounds without assumptions on P(X,Y)!
- c^ does not have to be a minimizer, it can be an approximator of the training data.
- Confidence (delta) and precision (epsilon) are limited.
- Global minimization on training data is a sin as it gives an atypical result.
- VC Inequality
  -
  - Pointwise convergence is enough for the c-star, as it has not seen any data and therefore will not fluctuate as n goes to infinity.
  - Bounding it by the supremum is wasteful whereas bounding ERM by the supremum is a necessity.
  - We need stronger guarantees than pointwise convergence for the ERM. We need uniform convergence, which gives rise to entropy.
  - As long as you have a selection criteria guided by the cost function this applies — whether you optimize or approximate.
  - **Intro. to ML**: For learning via empirical risk minimization to be successful, need uniform convergence.
  - Bernstein bounds limit the variance rather than support of the random variables (e.g. which may come useful for Gaussians). Hoeffding's bound becomes meaningless as the support gets very large.
  - The Hoeffding bound for binomial random variables is also called Chernoff or Okamoto bound.
  - n can be seen as the information source and N as the information sink. (n is the effective number of independent samples)
  - Counting the number of parameters to index concepts could be a wrong way to measure the cardinality of some C. Such type of bounds where we assume nothing about the probability distribution may be too conservative beacuse we are immunized against all overfitting events.
- The case of a finite hypothesis class can be generalized in two ways:

    - A hypothesis class with infinite cardinality is represented by finitely many hypotheses which yield different classifications on the data. (Quantization of the hypothesis space - fingering argument)
    - Measure the Vapnik-Chervonenkis (VC) dimension of a set of functions and select a hypothesis class H with $dimVC(H) < \infty$.
- The VC dimension of a model f is the maximum number of points that can be arranged so that f shatters them.

12 - Nonparametric Bayesian Methods

- Beta distribution interpretation: probability of a Bernoulli process after observing a-1 successes and b-1 failures
- Dirichlet distribution: multivariate generalization of the beta distribution
- Finite Gaussian Mixture Model, issues with selecting K, latent clusters, infinite K
- Fundamentals
    - **Wikipedia**: Dirichlet distribution is the conjugate prior of the categorical distribution and multinomial distribution.
    - **Wikipedia**: In probability theory and statistics, a sequence or collection of random variables is independent and identically distributed (i.i.d. or iid or IID) if each random variable has the same probability distribution as the others and all are mutually independent.
    - **Wikipedia**: In statistics, an exchangeable sequence of random variables (also sometimes interchangeable) is a sequence X_1, X_2, X_3, ... (which may be finitely or infinitely long) whose joint probability distribution does not change when the positions in the sequence in which finitely many of them appear is altered. This provides a useful generalization — for example, sampling without replacement is not independent, but is exchangeable — and is widely used in Bayesian statistics.
    - **Wikipedia**: A sequence of random variables that are independent and identically-distributed (i.i.d.), conditional on some underlying distributional form is exchangeable.
    - **Wikipedia**: de Finetti's theorem states that exchangeable observations are conditionally independent relative to some latent variable. The extended versions of the theorem show that in any infinite sequence of exchangeable random variables, the random variables are conditionally independent and identically-distributed, given the underlying distributional form.
    - **Wikipedia**: A way of putting this is that de Finetti's theorem characterizes exchangeable sequences as mixtures of i.i.d. sequences — while an exchangeable sequence need not itself be unconditionally i.i.d., it can be expressed as a mixture of underlying i.i.d. sequences. Note that not all finite exchangeable sequences are mixtures of i.i.d. To see this, consider sampling without replacement from a finite set until no elements are left. The resulting sequence is exchangeable, but not a mixture of i.i.d. Indeed, conditioned on all other elements in the sequence, the remaining element is known.
    - Exchangeable: order and labeling independent, any permutation should lead to the same probability distribution
    - i.i.d. is a more 'brutal' assumption on data than exchangeability
- Dirichlet process: Chinese restaurant process, Hoppe urn, stick-breaking process, Griffiths-Engen-McCloskey (GEM) distribution, rich-get-richer effect (preferential attachment)
    - Expected cluster count is the property of the DP as a prior (without seeing the data)
    - Polya urn is a finite mixture analogy, Hoppe urn is an infinite mixture analogy

    - Dirichlet process is the latent random variable in De Finetti's theorem for Hoppe urn / CRP
    - de Finetti tells us what is essential: we choose DP as it has the exchangeability property
- DP Generative Model
- Fitting
    - By exchangeability, we can change the assignment of the element without influencing other assignments
    - If the model does not ideally capture reality, we will always end up with infinite number of clusters
    - Model order selection problem: how to select K?
    - Not easy to fit these models, no analytical formulas are found, no closed form solutions for the estimation equations
    - To fit the Dirichlet Process Mixture Model (DPMM) we use a technique called Gibbs Sampling (this is why we need exchangeability)
    - Too much components creates a superfluous effect, too less components introduces a bias
- Latent Dirichlet Allocation
    - So far every point belonged to one and only one cluster. What if the data is generated from a multivariate distribution?

13 - Deep Generative Modeling

- Credit assignment problem: Which parameter to blame for a mismatch in your result
    - It is simple in perceptrons, handled by error backpropagation in multi-layer perceptrons (MLPs)
- Backpropagation
    - Learning rule is local.
    - Learning rule lacks biological plausibility