Mid-term Review

2020年10月28日 8:17

History

Golden Years
1st AI winter: search space explosion, reasoning problem
AI boom: expert system, neural network
2nd AI winter: expert system, nn fail
development

Basic Concepts

Training/test/validation set loss function

- validation: used for estimation of the model **while tuning the hyperparameters**

Empirical loss, population loss

- population loss: ultimate goal, but we cannot see
- optimize empirical loss, hope to generalize to population loss

Optimization vs generalization

- Memorization function (an example that optimization does not give generalization)

cross validation
generate dataset (labeling pipeline/recapture)

overfit vs underfit
classical (complex network easily overfit --> use regularization)
modern (although possible overfit almost never happen, implicit regularization)

unsupervised learning
semi-supervised learning

- unlabeled data also help optimization
- but not always! assumptions are needed:
- continuity assumption(points that are closer tend to share same labels)
- manifold assumption(high dim data live approximately on the lower dim space)

Gradient Descent

zeroth order, first order, second order

- zero-order (hyper parameter)
- first-order GD
- second order (computing Hessian matrix is slow, and we don't need so much accuracy)

smoothness, convexity, strong convexity

- smooth: $f(y) - f(x) - \langle \nabla f(x), (y - x) \rangle \leq \frac{L}{2} \left|\left|y - x\right|\right|^2$ OR $\lambda_{max}(\nabla^2 f(x)) \leq L$
- strong convexity: $f(y) - f(x) - \langle \nabla f(x), (y - x) \rangle \geq \frac{\mu}{2} \left|\left|y - x\right|\right|^2$ OR $\lambda_{min}(\nabla^2 f(x)) \geq \mu$
- saddle points

smoothness → mono-decrease
smoothness + convexity + GD → 1/T convergence rate (telescoping)
smoothness + strong convexity + GD → linear convergence rate

- Limitation of GD: local optimum
- SGD: the advantages of randomness, also faster
- SVRG analysis (improve convergence rate)
- SVRG not practical: converging too fast may not be good
- epoch means going through the dataset once, which means plenty of iterations (converge before traverse?)

- non-convex analysis (hw2: converge to a stationary point)

Linear regression
- perceptron algorithm (converge if the data is perfectly seperable)
- logistic regression
- cross entropy for probabilities (scale of GD automatically fixed)

Regularization
- Ridge(L2) hard limit on $||w||$, relaxation use L2 loss
- Ridge intuition: normal GD + weight decay
- Note: **weight decay $\neq$ L2 regularization!**
- Lasso(L1) find the important features from a large number of them (L0→L1)
- Lasso intuition: normal GD + pull to zero

Compressed Sensing
- Design a measurement matrix A to extract the desired features
- RIP property: acting on s-sparse vector does not change the length too much, guarantees recovery
- Recovery: use L1 to approximate L0

SVM
- margin: the minimum
- hard margin: perfect if $y_i w^T x_i \geq 1$ for all i
- soft margin: slack variables, hinge loss

kernel method:
- map X to a high dimensional space $\phi(X)$, consider the dual problem, only need to compute:
- $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$
- $a_i \neq 0$ support vectors
- example: quadratic kernel

Generalization theory
no free lunch theorem (no universal learner)
- proof intuition:
- $T = 2^{2m}$ different functions,
- Lower bound the expectation of population loss:
- k different training set sequences
- pairing the T functions: for a sample v_r not in training set, one predicts 0 and another predicts 1

Error decomposition
- approximation error: the best we can get in hypothesis class
- estimation error: increases with $|\mathrm{H}|$ and decrease with m

ERM algorithm

- $ERM_H(S) \in argmin_{h \in H} L_S(h)$

PAC learnable hypothesis class H

- For sample size $m \geq m(\epsilon, \delta)$, satisfying realizability assumption, outputs h with $L_{D,f}(h) \leq \epsilon$

Agnostic PAC learning (Bayes optimal predicter)

- $L_D(h) \leq \min_{h' \in H} L_D(h') + \epsilon$

Rademacher Complexity

- bound for expectation of representative
- bound for ERM loss
- Contraction lemma: Lipschitz bound for $R(A)$
- Massart Lemma: log exp and use Jensons inequality