MAS473 Introduction to DL - Midterm Summary

Yujun Kim

October 2023

1 Linear Algebra

- A vector space V with a subspace U has an orthogonal decompsition

$$V = U \oplus U^{\perp}$$

- Let $U \subseteq \mathbb{R}^n$. B be the matrix whose columns form the basis of U. Then,

$$P_U(x) = B(B^T B)^{-1} B^T x$$

- $Tr(AB) = Tr(BA)$
- Trace is invariant to the change of basis.
- (Spectral Theorem) A is normal if and only if it has an orthogonal diagonalization. In particular, every symmetric matrices are orthongally diagonalizable.
- A matrix A is PSD if and only if all its eigenvalues are nonnegative.
- To find SVD of a matrix A, find eigen decomposition of $A^T A$ or AA^T.

2 Probability Theory

2.1 Estimators

- A function P on a σ-algebra on Ω is a probability measure if (i) All its values lies on [0, 1], (ii) $P(\Omega) = 1$, (iii) Value on countable disjoint union is the countable sum of values on each.
- **(Bayes' Formula - Discrete)** If $\cup E_i = \Omega$,

$$P(E_i|F) = \frac{P(F|E_i)P(E_i)}{\sum_i P(F|E_i)P(E_i)}$$

- **(Maximum likelihood estimator)** Given $X_1, \cdots, X_n$ given from a distribution parameterized by θ, MLE maximize the joint distribution.

$$\theta_{MLE} = \text{argmax } f(x_1, \cdots, x_n|\theta)$$

- **(Bayesian Inferene)** Given a prior f_θ, we can compute posterior $f(\theta|x_1, \cdots, x_n)$ using Bayes's theorem. Bayes' estimator minimize MSE for the posterior distribution. It is given by
$$\theta_{Bayes} = \int \theta f(\theta|x_1, \cdots, x_n) d\theta$$
- Given a prior, **MAP Estimator** maximize the posteroir.
$$\theta_{MAP} = \text{argmax } f(\theta|x_1, \cdots, x_n)$$

2.2 KL Divergence and Entropy

- **(KL Divergence)** $KL(p||q) = \mathbb{E}_{p(X)}\left[log\left(\frac{p(X)}{q(X)}\right)\right] = \int p(x) log\left(\frac{p(x)}{q(x)}\right) dx$
- Using $-\log$ is convex and apply Jensen, $KL(p||q) \geq 0$.
- KL divergence is not symmetric, with counter example given by two distinct Bernoulli random variable.
- **(Entropy)** $H(X) = -\mathbb{E}_{p(X)}[log(p(X)]$
- **(Conditional Entropy)** $H(Y|X) = -\mathbb{E}_{p(x,y)}[logp(Y|X)]$
- **(Mutual Information)** $I(X;Y) = KL(p(x,y)||p(x)p(y)) = -\mathbb{E}_{p(x,y)}\left[log\left(\frac{p(x,y)}{p(x)p(y)}\right)\right]$
- **(Cross Entropy)** $H_p(q) = -\mathbb{E}_{p(x)}[log(q(x))]$
- $KL(p||q) = H_p(q) - H(p)$
- Minimizing cross entropy $Hp(q_\theta)$ in terms of θ is finding MLE.

(Question 1) Prove that discrete RV having at most n different values has entropy at least $log(n)$. When is this achieved? Can we extend this to the continuous case?

(Question 2) As a extension of written HW problem 1, (i) Show that right singular vector corresponding to largest singular value maximize $||Ax||_2$ over all unit vector x. (ii) Show second singular vector maximize $||Ax||_2$ over all unit vector x that is orthogonal to the first singular singular vector.

3 Optimization

- **(Lagrangian Dual)** The Lagrangian dual of the primal problem
$$\min_{x\in\mathbb{R}^n} f(x) \text{ subject to } g_i(x) \leq 0, \forall i = 1, \cdots, m$$
is given by
$$\max_{\lambda\in\mathbb{R}^n} \{\mathcal{D}(\lambda) \equiv \min_{x\in\mathbb{R}^n} \mathcal{L}(x,\lambda)\} \text{ subject to } \lambda \geq 0$$
where the Lagrangian function is given by $\mathcal{L}(x,\lambda) = f(x) + \sum_{i=1}^m \lambda_i g_i(x)$

- **(Weak Duality)** Optimal primal p^* and optimal dual d^* satisfies $d^* \leq p^*$. (proof) $\mathcal{D}(\lambda) \leq \mathcal{L}(x, \lambda) \leq f(x)$ for all feasible x, λ. Take maximum over feasible λ and minimum over feasible x

- Why we consider dual? Dual problem is convex. Dual problem has dimension different from primal(sometimes beneficial). Dual problem has simple constraint set.

- **(Separating Hyperplane theorem)** Nonempty disjoint convex set has a separating hyperplane. (proof) First consider the case two sets are compact. In this case, we can use EVT to prove the statement. Now in the case one is closed and one is compact, the distance is actually achieved by some point(and thus the distance is positive). For general case, consider closure of each set and take intersection with large enough ball to make one compact.

- **(Supporting Hyperplane Theorem)** A nonempty convex set has a supporting hyperplane at its boundary. (proof) A direct corollary of separating hyperplane with the given convex set and a one point set formed by a boundary point.

- **(Characterization of Compact Function)** TFAE for twice differentiable f: (i)f is convex. (ii)f is greater or equal to its tangent plane at every point. (iii)$\nabla^2 f \succeq 0$

- **GD implicitly solve regularized regression** in the overparameterized setting under a condition on initial point. Recall that on underparametrized setting, we solved

$$A^T(Ax - b) = 0$$

to achieve $Ax = A(A^TA)^{-1}A^Tb$. For overparametrized setting, A is a fat matrix so that A^TA is not invertible. Instead $x = A^T(AA^T)^{-1}b$ solves above equation, and any other solution is of the form $x + v$ where $v \in Null(A^TA) = Null(A)$. Hence, given x is the unique regularized solution. If $x_0 \in R(A^T)$, then GD iterates lies on the column space of A^T. Using $Null(A) \perp R(A^T)$, we see that GD converges to x

4 NN and Backpropagation

- Automatic Differentiation. Forward propagation for computation of variables with bacakpropagation for computation of derivatives.

5 Regression

- MLE of Gaussian distribution for the linear regression problem solves least square error problem.

- Kernel regression are used for nonlinear estimation.

- MAP estimator of Gaussian distribution for the kernel regression problem with a Gaussian prior solves regularized problem. By Bayes' theorem,

$$p(\theta|X, Y) \propto p(X, Y|\theta)p(\theta)$$

Also, $\frac{p(X,Y|\theta)}{p(Y|X,\theta)} = \frac{p(X,Y,\theta)p(X,\theta)}{p(X,Y,\theta)p(\theta)} = p(X|\theta)$ which is not related to θ. Thus,

$$p(\theta|X,Y) \propto p(Y|X,\theta)p(\theta)$$

, where proportional terms hide terms not related to θ. Thus, negative log-posteroir is

$$\frac{1}{2\sigma^2}||y - \Phi\theta||^2 + \frac{1}{2b^2}||\theta||^2$$

plus some constant term. Minimizing this solves l_2 regularized problem:

$$\theta = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T y$$

- **(Excess error is Bias+Variance+Irreducible error)** Let $y = f(x) + \epsilon$ for mean zero noise ϵ. Let $\hat{f}$ be the estimator of f.

$$\mathbb{E}[y^2] = f^2 + \sigma^2$$
$$\mathbb{E}[y\hat{f}] = f\mathbb{E}[\hat{f}]$$
$$\mathbb{E}[\hat{f}^2] = Var(\hat{f}) + \mathbb{E}[\hat{f}]^2$$

Thus, $\mathbb{E}[(y - \hat{f})^2] = (f - \mathbb{E}[\hat{f}])^2 + Var(\hat{f}) + \sigma^2$ If excess error is fixed, bias-variance has a trade off.

6 Classification

6.1 Binary Classification

Classification is based on logistic regression. $\sigma(z) = \frac{1}{1+\exp^{-z}}$ is strictly increasing and lies between 0 and 1. We model probability of x being positively labeled as $\sigma(x^T\theta)$ where x is extended by one more dimension to include a element 1. Finding MLE is a convex problem but does not have a closed form solution.

6.2 Multi Class Classification

We have two options

-
-

7 Support Vector Machine(SVM)

7.1 Hard Margin SVM

7.2 Soft Margin SVM

Convex surrogate loss - derivation from hinge loss.

7.3 Dual Soft Margin SVM

Not in examp