Midterm Review

Haojun Li

May 9, 2019

Table of Contents

Table of Contents

Kernel Recap

- Core idea: reparametrize parameter θ as a linear combination of featurized vectors.
- $\theta = \sum_{i=1}^{n} \beta_i \phi(\vec{x}^{(i)})$
- A kernel matrix K is a valid (mercer) kernel function if it is a PSD. Suppose $K(x^{(i)}, x^{(j)}) = \mathbf{K}_{ij}$
- If $\mathbf{K} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ then is K a valid kernel function?
- If $\mathbf{K} = \begin{bmatrix} 3 & 5 \\ 5 & 3 \end{bmatrix}$ then is K a valid kernel function?

Kernel Question (Midterm FA18)

Let us attempt to kernelize k-means algorithm. Consider the following update formula:

$$\text{step 1:} \quad c^{(i)[t+1]} := \arg\min_j ||x^{(i)} - \mu_j^{[t]}||^2$$

$$\text{step 2:} \quad \mu_j^{[t+1]} := \frac{\sum_{i=1}^m 1\{c^{(i)[t+1]} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)[t+1]} = j\}}$$

We seek to combine this into a single formula while also applying the kernel trick to allow infinite dimensional features. Complete the derivation below:

$$c^{(i)[t+1]} := \arg\min_j ||\phi(x^{(i)}) - \mu_j||^2$$

$$:=$$

Table of Contents

GLM Recap

- Exponential family is $P(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$
- Universal GLM update formula: $\theta := \theta + \alpha(y - h(x))x$
- At test time, we want to make a prediction: $h(x) = E[y|x] = a'(\eta)$
- We parametrize η as $\eta = \theta^T x$. (Or, if η is vector-valued, then $\eta_i = \theta_i^T x$)
- Assume that we have an exponential distribution

$$p(y;\lambda) = \lambda e^{-\lambda y}$$

What is $b(y)$, η, $T(y)$ and $a(\eta)$?

Generative Algorithm Recap

- We make a strong assumption about how our data and labels are generated.
  - Discriminative Algorithms - models $p(y|x;\theta)$ directly
  - Generative Algorithms - models $p(x,y;\phi) = p(x|y;\phi)p(y;\phi)$
- Basic formulation: Model the label as $p(y)$; model the data as $p(x|y)$

GLM + Generative Algorithm Question (Fall 2017 Midterm 3a)

- We examine a Generalized Discriminant Analysis (Generalized GDA)
- We model $p(y = 1; \phi) = \phi$ where $y \sim \text{Bernoulli}(\phi)$.
- We model $p(x|y = k; \eta_k) = b(x) \exp(\eta_k^T T(x) - a(\eta_k))$ where $T(x) = x$ and $k \in \{0, 1\}$ denotes the class label.
- Show that the decision boundary is linear by showing the posterior distribution is

$$p(y|x; \eta_0, \eta_1, \phi) = \frac{1}{1 + \exp(-(\theta_0 + \theta_1^\top x))}$$

Table of Contents

Learning Theory

- Bias-Variance trade off: Test MSE $= \sigma^2 + (\text{Bias}(\hat{f}(x)))^2 + Var(\hat{f}(x))$
- High σ means noisy data, nothing much we can do here so we are screwed if the training data is too noisy.
- High Bias is correlated with underfitting, your model is unable to correctly model the distribution.
- High Variance is correlated with overfitting, your model is too able to model the distribution.

Learning Theory (Supervised) (Fall 2018 Midterm 1b)

For the next set of questions, please indicate the effect on Bias and Variance (increase? Decrease? Uncertain?)
Assume that we are doing ordinary least squared regression on some data and labels

- Replace the value of your last feature by a random number for every example of your dataset.
- Map $x \in \mathbb{R}^n$ to $\phi(x) \in \mathbb{R}^d$ where $d > n$, $\phi(x)_i = x_i$ for $1 \leq i \leq n$, and $\phi(x)_i = f_i(x)$ for $n < i \leq d$ where each f_i is some scalar valued, deterministic function over the initial set of features. None of the f_i's are just constant values that ignore the input.
- Stop your optimization algorithm early.