SML lecture notes

Statistical Machine Learning

These are the notes for Statistical Machine Learning. The first lecture had a slide which detected health of a pig using it's picture. (Future prospects: Hiring/Admissions LOL). Anyways, I am using Obsidian and this is an amazing markdown editor! It has a lot of community plugins. Cool, let's get started!

Index

Lecture sigmoid(infinity)

Below is a big overview (But I already know everything here lol so nothing "new")

Classification

Predicting a discrete random variable Y from another random variable X.

- Consider data $(X_1, Y_1), \dots, (X_n, Y_n)$ where $X_i = (X_{i1}, X_{i2}, \dots, X_{id}) \in \mathcal{X} \subset$ $\mathbb{R}^d$
  is a d-dimensinoal vector and Y_i takes values in some finite set $\mathcal{Y}$. A **classification rule** is a function $h : \mathcal{X} \to \mathcal{Y}$. When we observe a new X we can predict Y to be $h(X)$.
- $Y = \{0, 1\}$ binary classification , rest maybe named as multiclass classification

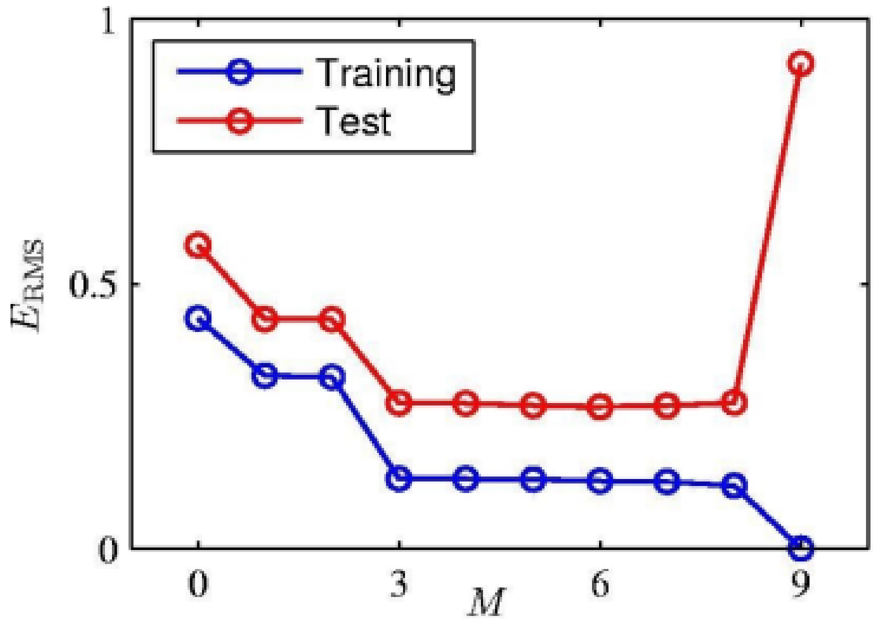
Loss Function

Say $y(x, \vec{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$

- $E(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2$: Sum of squares error function

Over-fitting

When training loss function is low but on testing it becomes high.



Regularization

This is just adding a term in the loss function to penalize when the magnitude of $\vec{w}$ is high. There is Lasso and Ridge regression (L1, L2). Lasso has sum of absolute values instead of sum of magnitude. Infact you may define your very own lol.

- $\widetilde{E}(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 + \frac{\lambda}{2}||w||^2$

Remaining class on reviing probability (class 10th level lmao)

Reference books

- Hastie, Tibshirani, Friedman Elements of Statistical Learning
- Murphy Machine Learning: a Probabilistic Perspective
- Duda: Pattern Classification

Evaluation

- Assignment (50%) - 5, This is pretty pog, I love the course ig
- Quiz (20%) - 3
- Midsem (15%)
- Endsem (15%)
- All mandatory

Grade cutoffs

- 91-100 A/A+ : Stupid course smh, hate the course (unless jsksksks)
- 81-90 A-
- 71-80 B

Further Reading

- Theoretical : AISTATS, ICML, JMLR, NeurIPS
- Systems+Theory: CVPR, ICCV, ECCV, AAAI, IEEE Transactions

L2 (Regularisation)

Unsupervised learning

Only data, no labels. Example PCA (dim reducition), K-means *clustering*

Looks like nothing was done here apart from revising probability lol.

PSD: Positive semi definite: Hermitian matrix with all eigenvalues positive.

Hermitian matrix when $A = \overline{A^T}$. (complex nos.)

Lecture tHr33

More revision.

Covariance

$$cov[x, y] = \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y^T - \mathbb{E}[y^T]\}]$$

so we define it for only one variable X, $cov(X) = \frac{1}{N-1}\sum_{i=1}^{N}[X_i - \mu_X][X_i - \mu_X]^T$ where $\mu_X = \mathbb{E}[x] \in \mathbb{R}^{d\times 1}$

The Gaussian Distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Looks like everyone is a fan of `$\mathcal$` Here, $\mathbb{E}[x] = \mu, var[x] = \sigma^2$

But in this non-binary world there are a lot of things. Presenting multivariate Gaussian (duh)

$$\mathcal{N}(x|\mu, \Sigma^2) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{\overbrace{(x-\mu)^T}^{1\times d} \Sigma^{-1} (x-\mu)}{2}}$$

where obviously $\Sigma = cov(X)$

When $d > N, \Sigma$ is not a full rank matrix (max N). So Σ^{-1} PSD.

- $r^2 = (x-\mu)^T \Sigma{-1}(x-\mu)$ is called a Mahalanobis distance from x to μ. Imagine.
- volume of hyperellipsoid corresponding to a Mahalanobis distance r $V = V_d |\Sigma|^{\frac{1}{2}} r^d$ where V_d is the volume of a d-dimensional unit-hemisphere.
- Higher the determinant for a fixed r and d, higher the scatter. For covariance matrices of independent variables, the determinant is large and thus scatter is more.

Idk what's happening anymore lol.

Now we will do **Bayesian Decision Theory**

SML/Lecture 4 SML/Lecture 5 SML/Lecture 6