CS 195f Midterm Review

Jason Pacheco
Soumya Ghosh

Overview

	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization Discrete Output y $\in \{1, \ldots, K\}$	clustering
Continuous	regression Continuous Output $y \in R$	dimensionality reduction

Generative vs Discriminative

- Generative methods model data . Discriminative models model boundaries between classes.
- Models are trained differently
  - Generative models maximize the joint likelihood
    - $\sum_{i=1}^{N} \log p(x_i, y_i|\theta)$
  - Discriminative models maximize the conditional likelihood
    - $\sum_{i=1}^{N} \log(p(y_i|x_i, \theta)$

K-NN classifier

- “looks at” k points in the training set nearest to the test data instance.
- $p(y = c|x, D_t K) = \frac{1}{K} \sum_{i \in N_k(x, D_t)} I(y_i = c)$
- Defines a neighborhood $\mathrm{N_{k(x,D}}_t)$ with a distance metric $d(x, x^*)$
- Non parametric – number of parameters can grow with data.

Problems?

- What happens with increasing dimensionality?

- What happens with increasing data?

Discriminant analysis

- Continuous features
- Each class is fitted with a Gaussian Distribution
- MLE estimate of the mean ?
- MLE estimate of the covariance?
- Generative or Discriminative?

Logistic Regression Classifier

- Parametric model with parameters *w*

- Likelihood:
  - $p(y_n|x_n, w) = \text{Bernoulli}(\text{y}|\sigma(w^T x))$
  - $\sigma(w^T x) = \frac{1}{\{1+\exp(-w^T x)\}}$
- Discriminative or Generative?

Logistic Form

- $P(y_n = 1|x_n, w) = \frac{\exp(w^T x)}{1+\exp(w^T x)}$
- $P(y_n = 0|x_n, w) = 1 - \frac{\exp(w^T x)}{1+\exp(w^T x)}$

  $= \frac{1}{1+\exp(w^T x)}$
- Why this particular form?
- LOR(x) =

  $\log \frac{p(y_n=1|x_n,w)}{p(y_n=0|x_n,w)} = \log\left(\exp(\mathrm{w^T x})\right) = \mathrm{w^T x}$
- Linear Decision Boundary

Logistic Regression …

- Predict
  - $y_n = 1 \; if \; p(y_n = 1|x_n, w) > p(y_n = 0|x_n, w)$
  - $w^T x_n > 0$

- Let's say for a test point x^*, I have
  - $w^T x^* = \; 10 + 100 x_i^* \; - 200 x_j^* + 0 \; x_k^*$

Logistic Regression - Nonlinearities

- Decision boundary is linear in data space.

- Can add more flexibility
  - by applying a feature transform or basis function expansion $\phi(x)$
  - Decision surface now is $w^T \phi(x)$
    - Linear in $\phi(x)$ but potentially non linear in x

Multiclass Extension

- $p(y_n|x_n, w) = \mathrm{Cat}(\mathrm{y_n}|S(W^T x))$

- $S_c(W^T x) = \frac{\exp(w_c^T x)}{\sum_k \exp(w_k^T x)}$

Naïve Bayes Classifier

- Naïve Bayes assumes conditional independence amongst features given class labels.
  - $x_{n_i} \perp x_{n_j} \mid y \,; j \neq i$
- Model
  - $p(X, Y|\theta) = \prod p(y_n|\theta) \prod_j^M p(x_{nj}, |y_n, \theta)$
  - θ = Model Parameters
- Generative or Discriminative?

Regression

- Continuous output $\mathbf{y} \in \mathbb{R}^M$

Linear Regression

- Output real-valued so we replace Bernoulli with Gaussian,

$$p(y|x,\theta) = N(y|w^T\phi(x),\sigma^2)$$

- Equivalent to,

$$y = w^T\phi(x) + \varepsilon, \qquad \text{where } \varepsilon \sim N(0,\sigma^2)$$

Regression

Linear Regression (cont'd)

- ML estimate w_{ML} yields *least squares* solution

$$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y}$$

- Where Φ is our *model matrix*

- Solution is unique global optimum

Decision Theory

Goal: compute decision procedure to minimize expected loss $\delta : \mathcal{X} \rightarrow \mathcal{A}$

- In Bayesian approach we minimize *posterior expected loss*

$$\rho(\mathbf{a}|\mathbf{x}, \pi) := \mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{x},\pi)} \left[L(\boldsymbol{\theta}, \mathbf{a}) \right] = \int_{\Theta} L(\boldsymbol{\theta}, \mathbf{a}) p(\boldsymbol{\theta}|\mathbf{x}, \pi) d\theta$$

Decision Theory

Loss Functions:

- We are free to choose any loss function
- MAP estimate minimizes 0-1 loss:

$$L(\theta, a) = \mathbb{I}(\theta \neq a) = \begin{cases} 0 & \text{if } a = \theta \\ 1 & \text{if } a \neq \theta \end{cases}$$

- Plugging into $\rho(a|\mathbf{x})$ yields

$$\rho(a|\mathbf{x}) = p(a \neq y|\mathbf{x}) = 1 - p(a=y|x)$$

$$y^*(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x}), \quad \textit{(MAP)}$$

Decision Theory

Loss Functions (cont'd):

- Posterior mean minimizes L2 loss,

$$L(\theta, a) = (\theta - a)^2$$

- Plug into expected loss,

$$\rho(a|\mathbf{x}) = \mathbb{E}\left[(\theta - a)^2|\mathbf{x}\right] = \mathbb{E}\left[\theta^2|\mathbf{x}\right] - 2a\mathbb{E}\left[\theta|\mathbf{x}\right] + a^2$$

$$\frac{\partial}{\partial a}\rho(a|\mathbf{x}) = -2\mathbb{E}\left[\theta|\mathbf{x}\right] + 2a = 0$$

$$\Rightarrow \; a = \mathbb{E}\left[\theta|\mathbf{x}\right] = \int \theta p(\theta|\mathbf{x}) d\theta$$

- Posterior median minimizes L1 loss

Decision Theory

Loss Functions (cont'd):

- Consider *loss matrix*

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	0	L_{FN}
$y=0$	L_{FP}	0

, L_{FN} – *False Negative*
, L_{FP} – *False Positive*

- Posterior expected loss is

$$\rho(\hat{y}=0|\mathbf{x}) = L_{FN} \times p(y=1|\mathbf{x})$$
$$\rho(\hat{y}=1|\mathbf{x}) = L_{FP} \times p(y=0|\mathbf{x})$$

- So we pick class $\hat{y}=1$ iff

$$\rho(\hat{y}=0|\mathbf{x}) > \rho(\hat{y}=1|\mathbf{x})$$
$$\frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} > \frac{L_{FP}}{L_{FN}} = \eta$$

Can trace out ROC curve
By varying η