Lecture 6: Applications

Xiugang Wu

University of Delaware

Fall 2019

Statistics and Machine Learning

Consider the problem of predicting $Y \in \mathcal{Y}$ when given the information of $X \in \mathcal{X}$. Here $X \in \mathcal{X}$ is called the feature and $Y \in \mathcal{Y}$ is called the label (or target). Note that the problem includes the special case when $\mathcal{X} = \emptyset$.

- Data generation mechanism: $(X, Y) \sim P$, with $P = P_X P_{XY}$

- Performance measure: Under loss function $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbf{R}_+$, the performance of predictor h is measured by the risk $L(h, P) = \mathrm{E}_P[\ell(Y, h(X))]$

- If P is known, the optimal predictor is given by the Bayes predictor

$$h^* = \arg\min_h \mathrm{E}_P[\ell(Y, h(X))]$$

- What if P is unknown and instead we have access to data $\{(X_i, Y_i)\}_{i=1}^n$ that are i.i.d. generated according to P?

Statistics and Machine Learning

Two approaches to the problem, which are generally known as the generative approach and the discriminative approach:

- Generative approach (statistical decision theory): Estimate the distribution P based on data $\{(X_i, Y_i)\}_{i=1}^n$ and then design the predictor; includes parametric and nonparametric estimation

- Discriminative approach (statistical learning theory): learn the predictor directly from data $\{(X_i, Y_i)\}_{i=1}^n$ without the intermediate step of estimating P; includes classification and regression

Outline

- Parametric Estimation
- Nonparametric Estimation
- Linear Regression and Logistic Regression
- Support Vector Machine

Outline

- Parametric Estimation
- Nonparametric Estimation
- Linear Regression and Logistic Regression
- Support Vector Machine

Parametric Estimation

distribution estimation problem: estimate probability density $p(y)$ of a random variable from observed data

parametric distribution estimation: choose from a family of densities $p(y; x)$, indexed by a parameter x

MLE (maximum likelihood estimation): $\text{maximize}_x \log p(y; x)$
- y is observed data
- $l(x) = \log p(y; x)$ is called log-likelihood function
- can add constraints $x \in C$ explicitly, or define $p(y; x) = 0$ for $x \notin C$
- a convex optimization problem if $\log p(y; x)$ is concave in x for fixed y

Linear Measurements with IID Noise

Linear measurement model: $y_i = a_i^T x + v_i, \quad i = 1, 2, \ldots, m$
- $x \in \mathbf{R}^n$ is vector of unknown parameters
- v_i is i.i.d. measurement noise, with density $p(z)$
- y_i is measurement: $y \in \mathbf{R}^m$ has density $p(y; x) = \prod_{i=1}^m p(y_i - a_i^T x)$

ML Estimate $\hat{x}_{\mathrm{ML}}$: any solution x of

$$\text{maximize} \quad l(x) = \sum_{i=1}^{m} \log p(y_i - a_i^T x)$$

Interpretation:
- estimate probability density $p(y)$ from observed data $y_1, y_2, \ldots, y_m$
- densities parameterized by x as $p(y; x)$; e.g., if noise is zero-mean, then problem becomes estimating the mean of y, which is of the form $(a_1^T x, \ldots, a_m^T x)$

Examples

- Gaussian noise $\mathcal{N}(0,\sigma^2)$: $p(z) = (2\pi\sigma^2)^{-1/2} e^{-z^2/(2\sigma^2)}$

$$L(x) = -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(a_i^T x - y_i)^2$$

ML estimate is LS solution

- Laplacian noise: $p(z) = (1/2a)e^{-|z|/a}$

$$L(x) = -m\log(2a) - \frac{1}{a}\frac{1}{2\sigma^2}\sum_{i=1}^{m}|a_i^T x - y_i|$$

ML estimate is solution to ℓ_1-norm minimization

MAP Estimation

Maximum a posteriori probability (MAP) estimation is a Bayesian version of maximum likelihood estimation, with a prior probability density on the underlying parameter x.

We assume that x (the vector to be estimated) and y (the observation) are random variables with a joint probability density $p(x, y) = p(x)p(y|x)$, where $p(x)$ is the prior density of x and $p(y|x)$ is the conditional density of y given x.

Given observation y, the Maximum a posteriori probability (MAP) estimation is to find x that maximizes the posterior density of x given y, i.e.

$$\begin{aligned}
\hat{x}_{\mathrm{MAP}} &= \arg\max_x p(x|y) \\
&= \arg\max_x p(x, y) \\
&= \arg\max_x p(y|x)p(x) \\
&= \arg\max_x (\log p(y|x) + \log p(x))
\end{aligned}$$

- MAP reduces to ML when x is uniformly distributed
- for any MLE problem with concave log-likelihood, we can add a prior density $p(x)$ that is log-concave, and the resulting MAP problem will be convex

Revisiting Linear Measurements with IID Noise

Linear measurement model: $y_i = a_i^T x + v_i, \quad i = 1, 2, \ldots, m$
- $x \in \mathbf{R}^n$ has prior density $p(x)$
- v_i is i.i.d. measurement noise, with density $p_z(z)$
- conditional density $p(y|x) = \prod_{i=1}^m p_z(y_i - a_i^T x)$

MAP Estimate $\hat{x}_{\text{MAP}}$ can be found by solving

$$\text{maximize} \quad \left(\sum_{i=1}^m \log p_z(y_i - a_i^T x) + \log p(x) \right)$$

- For example, if $p_z(z)$ is $\mathcal{N}(0, \sigma_z^2)$ and $p(x)$ is $\mathcal{N}(\bar{x}, \Sigma_x)$, then the MAP estimate can be found by solving the QP:

$$\text{minimize} \quad \left(\sum_{i=1}^m (y_i - a_i^T x)^2 + (x - \bar{x})^T \Sigma_x^{-1} (x - \bar{x}) \right)$$

Outline

- Parametric Estimation
- Nonparametric Estimation
- Linear Regression and Logistic Regression
- Support Vector Machine

Nonparametric Estimation

Consider a discrete random variable X that takes values in the finite set $\mathcal{X} = \{a_1, \dots, a_n\}$ and let p_i denote the probability of X being equal to a_i, i.e. $p_i = p_x(a_i)$. The nonparametric estimation problem is to estimate p from the probability simplex

$$\{p \mid p \succeq 0, \mathbf{1}^T p = 1\}$$

based on a combination of prior information and, possibly, observations.

- many types of prior information about p can be expressed as linear equality or inequality constraints of p, e.g., $\mathrm{E}[x] = \sum_{i=1}^n a_i \cdot p_i = 3.3$, $\mathrm{E}[x^2] = \sum_{i=1}^n a_i^2 \cdot p_i \geq 4$, $\mathrm{E}[f(x)] = \sum_{i=1}^n f(a_i) \cdot p_i \in [l, u]$, $\mathrm{Pr}(X \in C) = \sum_{a \in C} p(a) = 0.3$

- can also include prior constraints involving nonlinear functions of p, e.g., $\mathrm{var}(x) = \mathrm{E}[x^2] - \mathrm{E}[x]^2 = \sum_{i=1}^n a_i^2 \cdot p_i - (\sum_{i=1}^n a_i \cdot p_i)^2$; a lower bound on the variance of X can be expressed as a convex quadratic inequality on p

- As another example, the prior constraint $\mathrm{Pr}(X \in A | X \in B) \in [l, u]$ can be expressed as $c^T p / d^T p \in [l, u]$, i.e. $l d^T p \leq c^T p \leq u d^T p$

- In general, we can express the prior information about the distribution p as $p \in \mathcal{P}$. We assume that $\mathcal{P}$ can be described by a set of linear equalities and convex inequalities, including the basic constraints $p \succeq 0, \mathbf{1}^T p = 1$

Nonparametric Estimation

Maximum Likelihood Estimation: Suppose we observe N independent samples $x_1, \ldots, x_N$ from the distribution. Let k_i denote the number of these samples with value a_i, so that $k_1 + \cdots + k_n = N$. The log-likelihood function is then $l(x) = \sum_{i=1}^n k_i \log p_i$, which is a concave function of p. The ML estimate of p can be found by solving the convex problem

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^n k_i \log p_i \\ \text{subject to} & p \in \mathcal{P} \end{array}$$

Maximum Entropy Estimation: The maximum entropy distribution consistent with the prior assumptions can be found by solving the convex problem

$$\begin{array}{ll} \text{maximize} & -\sum_{i=1}^n p_i \log p_i \\ \text{subject to} & p \in \mathcal{P} \end{array}$$

where the objective function $-\sum_{i=1}^n p_i \log p_i$ is concave in p.

Outline

- Parametric Estimation
- Nonparametric Estimation
- Linear Regression and Logistic Regression
- Support Vector Machine

Linear Regression

Linear regression with squared-loss: learn linear predictor $h_a(x) = a^T x$ from training data $\{(x_i, y_i)\}_{i=1}^n$

- $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \mathbf{R}$, $\hat{\mathcal{Y}} = \mathbf{R}$
- $a \in \mathbf{R}^d$ is the parameter to be learned
- loss function $\ell(y, \hat{y}) = (y - \hat{y})^2$
- risk of h_a under distribution P: $L(h_a, P) = \mathrm{E}_P[(Y - a^T X)^2]$
- empirical risk of h_a: $\frac{1}{n}\sum_{i=1}^n (y_i - a^T x_i)^2 = L(h_a, \hat{P})$, where $\hat{P}$ denotes the empirical distribution of (X, Y)

Empirical Risk Minimization (ERM):

$$\text{minimize}_a \sum_{i=1}^n (y_i - a^T x_i)^2$$

which is an ordinary least-squares (OLS) problem

Regularization

Two types of regularization: constrained ERM and Penalized ERM, where constrained ERM explicitly constrains the complexity of the model, and penalized ERM penalizes models with high complexity

constrained ERM: e.g., linear regression with constraint on ℓ_1 or ℓ_2 norm

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n}(y_i - a^T x_i)^2 \\ \text{subject to} & \|a\|_1 \leq r \end{array}$$

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n}(y_i - a^T x_i)^2 \\ \text{subject to} & \|a\|_2^2 \leq r \end{array}$$

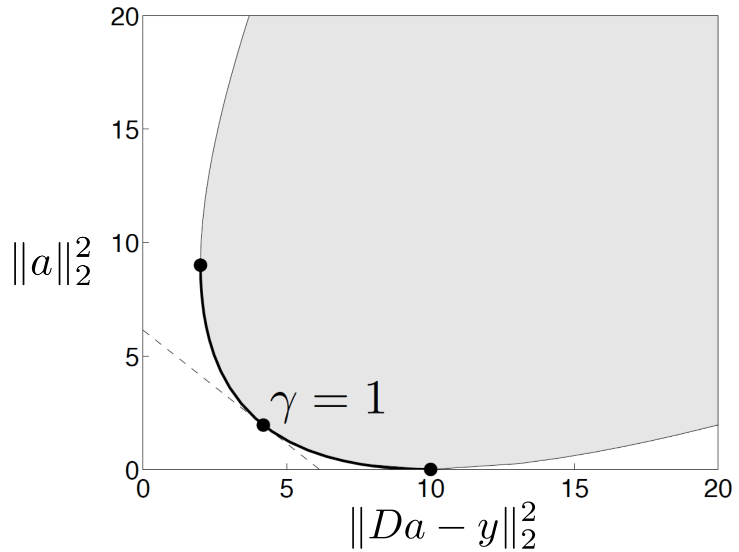
Penalized ERM: linear regression with regularizer of ℓ_1 or ℓ_2 norm

$$\text{minimize} \quad \sum_{i=1}^{n}(y_i - a^T x_i)^2 + \gamma\|a\|_1 \qquad \text{(LASSO Regression)}$$

$$\text{minimize} \quad \sum_{i=1}^{n}(y_i - a^T x_i)^2 + \gamma\|a\|_2^2 \qquad \text{(Ridge Regression)}$$

Multi-criterion Interpretation

$$\text{minimize (w.r.t. } \mathbf{R}_+^2) \quad (\|Da - y\|_2^2, \|a\|_2^2)$$



- example for $D \in \mathbf{R}^{100\times 10}$ with $D = [x_1^T; x_2^T; \dots, x_{100}^T]$; heavy line formed by Pareto optimal points

- to determine Pareto optimal points, take $\lambda = (1, \gamma)$ with $\gamma > 0$ and minimize

$$\|Da - y\|_2^2 + \gamma\|a\|_2^2$$

- for fixed γ, an OLS problem

In general, constrained ERM and penalized ERM and equivalent if criterion functions are all convex

Logistic Regression

Logistic regression: learn predictor $h_a(x) = \frac{1}{1+e^{-a^T x}}$ from training data $\{(x_i, y_i)\}_{i=1}^n$

- $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \{+1, -1\}$, $\hat{\mathcal{Y}} = [0, 1]$
- $\hat{y}$ is the predicted probability of the label of x being 1
- $a \in \mathbf{R}^d$ is the parameter to be learned
- loss function $\ell(y, h_a(x)) = \log(1 + e^{-ya^T x})$
- risk of h_a under distribution P: $L(h_a, P) = \mathrm{E}_P[\log(1 + e^{-Ya^T X})]$
- empirical risk of h_a: $\frac{1}{n}\sum_{i=1}^n \log(1 + e^{-y_i a^T x_i}) = L(h_a, \hat{P})$

Empirical Risk Minimization (ERM):

$$\text{minimize}_a \sum_{i=1}^n \log(1 + e^{-y_i a^T x_i})$$

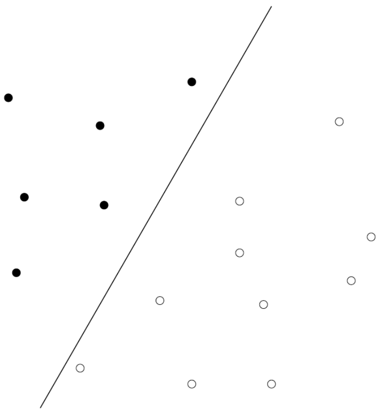
which is a convex optimization problem. Convexity of the optimization problem inherits from the convexity of the loss function for a given data point (x, y).

Outline

- Parametric Estimation
- Nonparametric Estimation
- Linear Regression and Logistic Regression
- Support Vector Machine

Classification

Binary classfication: learn halfspace predictor $h_a(x) = \mathrm{sgn}(a^T x)$
- $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \hat{\mathcal{Y}} = \{+1, -1\}$
- $a \in \mathbf{R}^d$ is the parameter to be learned
- loss function $\ell(y, h_a(x)) = I(y \neq \mathrm{sgn}(a^T x))$
- risk of h_a under distribution P: $L(h_a, P) = \mathrm{E}_P[I(Y \neq \mathrm{sgn}(a^T X))]$
- empirical risk of h_a: $\frac{1}{n}\sum_{i=1}^n I(y_i \neq \mathrm{sgn}(a^T x_i)) = L(h_a, \hat{P})$



Empirical Risk Minimization (ERM):

$$\text{minimize}_a \sum_{i=1}^{n} I(y_i \neq \mathrm{sgn}(a^T x_i))$$

- If data is linearly separable, then there exists some a_1 such that $y_i a_1^T x_i > 0,\ \forall i$.
- Let $a_2 = \frac{a_1}{\min_i y_i a_1^T x_i}$. Then we have $y_i a_2^T x_i \geq 1,\ \forall i$.
- Therefore, ERM is equivalent to the feasible problem:

$$\text{find } a \ \text{ subject to } \ y_i a^T x_i \geq 1,\ \forall i$$

- There are infinitely many ERM solutions. Which one should we pick?

Support Vector Machine

Support Vector Machine (SVM) seeks for an ERM hyperplane that separates the training set with the largest margin
- margin γ of a hyperplane with respect to a training set is the minimal Euclidean distance between a point in the training set and the hyperplane
- $\gamma = \min_i y_i a^T x_i / \|a\|$
- If we scale a such that $\min_i y_i a^T x_i = 1$, then $\gamma = 1/\|a\|$

Support Vector Machine (SVM):

$$\text{minimize } \|a\|^2 \quad \text{subject to } \; y_i a^T x_i \geq 1, \forall i$$