ECE 5424: Introduction to Machine Learning

Topics:
- Gaussians
- (Linear) Regression

Readings: Barber 8.4, 17.1, 17.2

Stefan Lee
Virginia Tech
Administrativia

• HW1
  – Due tomorrow night be 11:55pm

• Project Proposal
  – Due: 09/21, 11:55 pm
  – <=2pages, NIPS format
Recap of last time
Statistical Estimation

• Frequentist Tool
  • Maximum Likelihood

• Bayesian Tools
  • Maximum A Posteriori
  • Bayesian Estimation
MLE

\[ \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]

- \( D_1 = \{1,1,1,0,0,0\} \)
- \( D_2 = \{1,0,1,0,1,0\} \)

- A function of the data \( \phi(Y) \) is a sufficient statistic, if the following is true

\[ \sum_{i \in D_1} \phi(y_i) = \sum_{i \in D_2} \phi(y_i) \implies L(\theta; D_1) = L(\theta; D_2) \]
Beta prior distribution – $P(\theta)$

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1 - \theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

MAP for Beta distribution

\[ P(\theta \mid D) = \frac{\theta^{\beta_H + \alpha_H - 1}(1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- MAP: use most likely parameter:
  \[ \hat{\theta}_{MAP} = \arg\max_\theta P(\theta \mid D) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \]

- Beta prior equivalent to extra W/L matches
- As \( N \to \infty \), prior is “forgotten”
- But, for small sample size, prior is important!
Effect of Prior

- **Prior = Beta(2,2)**
  - $\theta_{\text{prior}} = 0.5$

- **Dataset = \{H\}**
  - $L(\theta) = \theta$
  - $\theta_{\text{MLE}} = 1$

- **Posterior = Beta(3,2)**
  - $\theta_{\text{MAP}} = (3-1)/(3+2-2) = 2/3$
Effect of Prior

Starting from different priors

\[ P(X=H) \]

\[ M = \#\text{samples} \]
Using Bayesian posterior

- Posterior distribution:

\[ P(\theta \mid \mathcal{D}) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- Bayesian inference:
  - No longer single parameter:

\[ E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) \, d\theta \]

  - Integral is often hard to compute
Bayesian learning for multinomial

- What if you have a k sided coin???
- Likelihood function if \textit{categorical}:
Simplex

\[ 0 \leq \theta_k \leq 1 \]

\[ \sum_{k=1}^{3} \theta_k = 1 \]
Bayesian learning for multinomial

- What if you have a k sided coin???
- Likelihood function if **categorical**:

  - **Conjugate** prior for multinomial is **Dirichlet**:

    \[
    \theta \sim \text{Dirichlet}(\beta_1, \beta_2, \ldots, \beta_k) \propto \prod_i \theta_i^{\beta_i-1}
    \]
Mean:  
\[ \mathbb{E}[\theta_i] = \frac{\beta_i}{\sum_j \beta_j} \]

Mode:  
\[ \hat{\theta}_i = \frac{\beta_i - 1}{\sum_j \beta_j - k} \]
Dirichlet Probability Densities

• Matlab Demo
  – Written by Iyad Obeid
Dirichlet Samples

Samples from Dir (alpha=0.1)

Samples from Dir (alpha=1)

Dir(\theta | 0.1, 0.1, 0.1, 0.1, 0.1)

Dir(\theta | 1.0, 1.0, 1.0, 1.0, 1.0)

Slide Credit: Erik Sudderth
Bayesian learning for multinomial

• What if you have a k sided coin???
• Likelihood function if categorical:

Conjugate prior for multinomial is Dirichlet:

\[ \theta \sim \text{Dirichlet}(\beta_1, \beta_2, \ldots, \beta_k) \propto \prod_i \theta_i^{\beta_i - 1} \]

• Observe \( n \) data points, \( n_i \) from assignment \( i \), posterior:

Homework 1!!!! 😊
Plan for Today

• Gaussians
  – PDF
  – MLE/MAP estimation of mean

• Regression
  – Linear Regression
  – Connections with Gaussians
Gaussians
What about continuous variables?

• Boss says: If I want to bet on continuous variables, like stock prices, what can you do for me?

• You say: Let me tell you about Gaussians…

\[ P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
Why Gaussians?

• Why does the entire world seem to always be telling you about Gaussian?
  – Central Limit Theorem!
Central Limit Theorem

• Simplest Form
  – $X_1, X_2, \ldots, X_N$ are IID random variables
  – Mean $\mu$, variance $\sigma^2$
  – Sample mean $S_N$ approaches Gaussian for large $N$

• Demo
  – http://www.stat.sc.edu/~west/javahtml/CLT.html
Curse of Dimensionality

• Consider: Sphere of radius 1 in d-dims

• Consider: an outer $\epsilon$-shell in this sphere

• What is $\frac{\text{shell volume}}{\text{sphere volume}}$?
Why Gaussians?

• Why does the entire world seem to always be harping on about Gaussians?
  – Central Limit Theorem!
  – They’re easy (and we like easy)
  – Closely related to squared loss (will see in regression)
  – Mixture of Gaussians are sufficient to approximate many distributions (will see it clustering)
Some properties of Gaussians

• Affine transformation
  – multiplying by scalar and adding a constant
  – $X \sim N(\mu, \sigma^2)$
  – $Y = aX + b \Rightarrow Y \sim N(a\mu + b, a^2\sigma^2)$

• Sum of Independent Gaussians
  – $X \sim N(\mu_X, \sigma_X^2)$
  – $Y \sim N(\mu_Y, \sigma_Y^2)$
  – $Z = X+Y \Rightarrow Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
Learning a Gaussian

• Collect a bunch of data
  – Hopefully, i.i.d. samples
  – e.g., exam scores

• Learn parameters
  – Mean
  – Variance

\[
P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]
**MLE for Gaussian**

- Prob. of i.i.d. samples $D=\{x_1, \ldots, x_N\}$:

$$P(D \mid \mu, \sigma) = \left(\frac{1}{\sigma \sqrt{2\pi}}\right)^N \prod_{i=1}^N e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- Log-likelihood of data:

$$\ln P(D \mid \mu, \sigma) = \ln \left[ \left(\frac{1}{\sigma \sqrt{2\pi}}\right)^N \prod_{i=1}^N e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right] = -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^N \frac{(x_i-\mu)^2}{2\sigma^2}$$
Your second learning algorithm: MLE for mean of a Gaussian

- What’s MLE for mean?

\[
\frac{d}{d\mu} \ln P(D \mid \mu, \sigma) = \frac{d}{d\mu} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \right]
\]
MLE for variance

- Again, set derivative to zero:

$$\frac{d}{d\sigma} \ln P(D | \mu, \sigma) = \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$= \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} \right] - \sum_{i=1}^{N} \frac{d}{d\sigma} \left[ \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$
Learning Gaussian parameters

- MLE:

\[
\hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i
\]

\[
\hat{\sigma}^2_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\]
Bayesian learning of Gaussian parameters

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Inverse Gamma or Wishart Distribution

- Prior for mean:

\[ P(\mu \mid \eta, \lambda) = \frac{1}{\lambda \sqrt{2\pi}} e^{-\frac{(\mu-\eta)^2}{2\lambda^2}} \]
MAP for mean of Gaussian

\[ P(\mu \mid \eta, \lambda) = \frac{1}{\lambda \sqrt{2\pi}} e^{-\frac{(\mu - \eta)^2}{2\lambda^2}} \]

\[ P(\mathcal{D} \mid \mu, \sigma) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^N \prod_{i=1}^{N} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \]

\[
\frac{d}{d\mu} \left[ \ln P(\mathcal{D} \mid \mu) P(\mu) \right] = \frac{d}{d\mu} \left[ \ln P(\mathcal{D} \mid \mu) + \ln P(\mu) \right]
\]
New Topic: Regression
1-NN for Regression

- Often bumpy (overfits)

Figure Credit: Andrew Moore
Linear fitting to data

- We want to fit a linear function to an observed set of points \( X = [x_1, \ldots, x_N] \) with associated labels \( Y = [y_1, \ldots, y_N] \).
  - Once we fit the function, we want to use it to predict the \( y \) for new \( x \).
Linear fitting to data

- We want to fit a linear function to an observed set of points $X = [x_1, \ldots, x_N]$ with associated labels $Y = [y_1, \ldots, y_N]$.
  - Once we fit the function, we want to use it to predict the $y$ for new $x$.
- Least squares (LSQ) fitting criterion: find the function that minimizes sum (or average) of square distances between actual $y$s in the training set and predicted ones.
Linear fitting to data

- We want to fit a linear function to an observed set of points \( X = [x_1, \ldots, x_N] \) with associated labels \( Y = [y_1, \ldots, y_N] \).
  - Once we fit the function, we want to use it to predict the \( y \) for new \( x \).

- Least squares (LSQ) fitting criterion: find the function that minimizes sum (or average) of square distances between actual \( y \)s in the training set and predicted ones.

The fitted line is used as a predictor.
Linear Regression

- Demo
  - http://hspm.sph.sc.edu/courses/J716/demos/LeastSquares/LeastSquaresDemo.html
Linear functions

- General form: \( f(x; w) = w_0 + w_1 x_1 + \ldots + w_d x_d \)
- 1D case (\( \mathcal{X} = \mathbb{R} \)): a line
- \( \mathcal{X} = \mathbb{R}^2 \): a plane
- \textit{Hyperplane} in general, \( d \)-D case.
Least squares: estimation

- We need to minimize w.r.t. \( w \)

\[
L(w, X) = L(w) = \frac{1}{N} \sum_{i=1}^{N} (y_i - w^T x_i)^2
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_{i1} - \ldots - w_d x_{id})^2
\]

- Necessary condition to minimize \( L \): derivatives w.r.t. \( w_0, w_1, \ldots, w_d \) must be zero.
Least squares in matrix form

\[ X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nd} \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \quad w = \begin{bmatrix} w_0 \\ \vdots \\ w_d \end{bmatrix}. \]

- Predictions: \( \hat{y} = Xw \), errors: \( y - Xw \), empirical loss:

\[
L(w, X) = \frac{1}{N} (y - Xw)^T (y - Xw)
\]