Leren — Homework 2
Tim Stolp 11848782

Chapter 4 & 5, Alpaydin

Deadline: 23:59 November 13th, 2018

This is the third week's assignment for Leren. This assignment covers chapter 4 & 5 of Alpaydin. Please take note of the following:

- You are expected to hand in your solutions in LaTeX;
- This problem set is an individual assignment;
- The deadline for this first assignment is Tuesday, November 13 at 23:59.

1 Background: Linearity of the expectation

Use the definition of the expectation to answer the following questions:

(a) For a random variable X,

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b \tag{1}$$

Show that this is the case.

E[aX + b] = $\sum_x (ax + b)p(x)$
$= \sum_x axp(x) + \sum_x bp(x)$
$= a\sum_x xp(x) + b\sum_x p(x)$
The definition of the expectation shows that $\mathbb{E}(X) = \sum_x xp(x)$ and $\sum_x p(x) = 1$, this results in:
$= a\mathbb{E}(X) + b$

(b) For random variables X and Y,

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \tag{2}$$

Show that this is the case.

$\mathbb{E}$[X + Y] = $\sum_i \sum_j (x_i + y_j)p(x_i, y_j)$
$\sum_i \sum_j x_i p(x_i, y_j) + \sum_i \sum_j y_i p(x_i, y_j)$
$= \mathbb{E}(X) + \mathbb{E}(Y)$

2 Chapter 4: Parametric Methods

2.1 Bias/Variance

(a) We have some model $g(x)$ that has a very high bias. What does that mean in terms of the training and validation errors?

High bias means that the model underfits the data. This results in a high trainingerror and an equally high validation error, because the classifier does not account for relevant information present in the training set.

(b) We also consider some other model, $h(x)$ that has a very high variance. What does that mean in terms of the training and validation errors?

High variance means that the model overfits the data. The training error is low, because of the overfitting. The validation error is high because of the overfitting the model does not generalize well.

(c) Suppose we use a massive dataset for training both models, which one is likely to give a better performance? Will the model that performs worse overfit or underfit?

A high variance is more likely to give a better performance. The model that performs worse will underfit.

2.2 MAP Estimation of Gaussian (normal) Density

In this exercise, we will lead you step-by-step to derive the Maximum A Posteriori (MAP) estimate of the mean of a Gaussian density. Recall that the MAP is given by $\theta_{MAP} = \arg\max_{\theta} p(\theta|X)$ (where in this case, the parameter θ is μ). Assume we are trying to fit a Gaussian to a dataset X with N data points x_i.

$$P(X|\mu,\sigma_0) = \frac{1}{(2\pi)^{N/2}\sigma_0^N} \exp\left\{-\frac{\sum_{i=1}^{N}(\mathrm{x}_i-\mu)^2}{2\sigma_0^2}\right\} \tag{3}$$

where we have placed a prior on μ:

$$P(\mu|m,\sigma_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(\mu-m)^2}{2\sigma_1^2}\right\} \tag{4}$$

(a) Write down the posterior for this model, assuming that σ_0, m and σ_1 are given.

$P(\mu|X,\sigma_0,\sigma_1,m) = \frac{P(X|\mu,\sigma_0)P(\mu|m,\sigma_1)}{P(X)}$

(b) Write down the log of the posterior and fill in Eq. (b) and Eq. (b). Since $P(x)$ does not depend on μ, you can follow Eq. 4.39 from the book and write it as a constant c.

$$\begin{aligned}
&log(P(\mu|X,\sigma_0,\sigma_1,m)) = log(\tfrac{P(X|\mu,\sigma_0)P(\mu|m,\sigma_1)}{P(X)})\\
&= logP(X|\mu,\sigma_0) + logP(\mu|m,\sigma_1) - logP(X)\\
&= log\left(\tfrac{1}{(2\pi)^{N/2}\sigma_0^N}\exp\left\{-\tfrac{\sum_{i=1}^{N}(\mathrm{x}_i-\mu)^2}{2\sigma_0^2}\right\}\right) + log\left(\tfrac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\tfrac{(\mu-m)^2}{2\sigma_1^2}\right\}\right) - c\\
&= log(1) - log((2\pi)^{N/2}\sigma_0^N) - \tfrac{\sum_{i=1}^{N}(\mathrm{x}_i-\mu)^2}{2\sigma_0^2} + log(1) - log(\sqrt{2\pi}\sigma_1) - \tfrac{(\mu-m)^2}{2\sigma_1^2} - c
\end{aligned}$$

$$= -log((2\pi)^{N/2}\sigma_0^N) - \frac{\sum_{i=1}^{N}(\mathrm{x}_i-\mu)^2}{2\sigma_0^2} - log(\sqrt{2\pi}\sigma_1) - \frac{(\mu-m)^2}{2\sigma_1^2} - c$$
$$= -Nlog((2\pi)^{1/2}\sigma_0) - \frac{1}{2\sigma_0^2} \cdot \sum_{i=1}^{N}(\mathrm{x}_i-\mu)^2 - log(\sqrt{2\pi}\sigma_1) - \frac{(\mu-m)^2}{2\sigma_1^2} - c$$

(c) Take the derivative of the log posterior wrt μ.

$$\frac{\partial}{\partial\mu}log(P(\mu|X,\sigma_0,\sigma_1,m))$$
$$= \frac{\partial}{\partial\mu}(-Nlog((2\pi)^{1/2}\sigma_0) - \frac{1}{2\sigma_0^2} \cdot \sum_{i=1}^{N}(\mathrm{x}_i-\mu)^2 - log(\sqrt{2\pi}\sigma_1) - \frac{(\mu-m)^2}{2\sigma_1^2} - c)$$
$$= -\frac{1}{2\sigma_0^2} \cdot \sum_{i=1}^{N} -2(\mathrm{x}_i-\mu) - \frac{2(\mu-m)}{2\sigma_1^2}$$
$$= \frac{1}{\sigma_0^2} \cdot \sum_{i=1}^{N}(\mathrm{x}_i-\mu) - \frac{(\mu-m)}{\sigma_1^2}$$

(d) To obtain the MAP, we have to find the point where the log posterior is maximized. To find this point, set the derivative to 0 and solve for μ.

$$\frac{\partial}{\partial\mu}log(P(\mu|X,\sigma_0,\sigma_1,m)) = 0$$
$$\frac{1}{\sigma_0^2}\sum_{i=1}^{N}(\mathrm{x}_i-\mu) - \frac{(\mu-m)}{\sigma_1^2} = 0$$
$$\frac{1}{\sigma_0^2}\sum_{i=1}^{N}(\mathrm{x}_i) - \frac{1}{\sigma_0^2}N\mu - \frac{(\mu-m)}{\sigma_1^2} = 0$$
$$\frac{1}{\sigma_0^2}\sum_{i=1}^{N}(\mathrm{x}_i) - \frac{1}{\sigma_0^2}N\mu - \frac{\mu}{\sigma_1^2} + \frac{m}{\sigma_1^2} = 0$$
$$\frac{1}{\sigma_0^2}N\mu + \mu \cdot \frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2}\sum_{i=1}^{N}(\mathrm{x}_i) + \frac{1}{\sigma_1^2} \cdot m$$
$$\mu(\frac{1}{\sigma_0^2}N + \frac{1}{\sigma_1^2}) = \frac{1}{\sigma_0^2}\sum_{i=1}^{N}(\mathrm{x}_i) + \frac{1}{\sigma_1^2} \cdot m$$
$$\mu = \frac{\frac{1}{\sigma_0^2}\sum_{i=1}^{N}(\mathrm{x}_i) + \frac{1}{\sigma_1^2}\cdot m}{\frac{1}{\sigma_0^2}N + \frac{1}{\sigma_1^2}}$$

2.3 Estimators

Show that the estimator for the mean of a Gaussian density, $m = \frac{\sum_t x^t}{N}$ is a consistent estimator for the true mean μ, i.e. that Var(m)→ 0 as $N \to \infty$, assuming that the data is i.i.d.
Hint: Var($\sum_i x_i$) = $\sum_i$ Var(x_i) for i.i.d.

$$Var(m) \to 0 \text{ as } N \to \infty$$
$$Var(\frac{\sum_t x^t}{N})$$
$$= \frac{1}{N^2}\sum_t Var(x^t)$$
$$= \frac{N\sigma^2}{N^2}$$
$$= \frac{\sigma^2}{N}$$
$$= \lim_{N\to\infty}\frac{\sigma^2}{N} = 0$$

3 Chapter 5: Multivariate Methods

3.1 Maximum Likelihood of a Multivariate Gaussian Density

Given a dataset $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N$ with N d-dimensional (iid) data points. We assume that $\boldsymbol{x}_i$ are drawn from a Multivariate Gaussian distribution such that the likelihood function can be expressed as $l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{X}) \equiv P(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$P(\boldsymbol{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\boldsymbol{\pi})^{N \cdot D/2}|\boldsymbol{\Sigma}|^{N/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right\} \tag{5}$$

Derive the maximum likelihood estimator for the mean of the Gaussian. To do so, take the following steps:

(1) Write down the expression for the log-likelihood $\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{X}) \equiv \log\, l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{X})$

$$log\left(\tfrac{1}{(2\boldsymbol{\pi})^{N \cdot D/2}|\boldsymbol{\Sigma}|^{N/2}} \exp\left\{ -\tfrac{1}{2} \textstyle\sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right\} \right)$$

$= \text{-} \log((2\boldsymbol{\pi})^{N \cdot D/2}|\boldsymbol{\Sigma}|^{N/2}) - \frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$

$= -\frac{N}{2} \cdot log((2\boldsymbol{\pi})^D |\boldsymbol{\Sigma}|) - \frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$

(2) Take the derivative of the log-likelihood wrt $\boldsymbol{\mu}$ (Hint: $\frac{\partial(\mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a})}{\partial \mathbf{a}} = 2\mathbf{a}^T \boldsymbol{\Sigma}$)

$\frac{\partial}{\partial \mu} log(l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{X})) = \frac{\partial}{\partial \mu}\left(-\frac{N}{2} \cdot log((2\boldsymbol{\pi})^D |\boldsymbol{\Sigma}|) - \frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right)$

$= -\frac{1}{2} \sum_{i=1}^{N} 2(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}$

$= -\sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}$

(3) Set the derivative to 0

$\frac{\partial}{\partial \mu} log(l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{X})) = 0$
$-\sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} = 0$
$\boldsymbol{\Sigma}^{-1} > 0$ so,
$\sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^T = 0$
$\sum_{i=1}^{N} (\boldsymbol{x}_i^T - \boldsymbol{\mu}^T) = 0$
$-N\boldsymbol{\mu}^T + \sum_{i=1}^{N} \boldsymbol{x}_i^T = 0$

(4) Solve for $\boldsymbol{\mu}$

$N\boldsymbol{\mu}^T = \sum_{i=1}^{N} \boldsymbol{x}_i^T$
$\boldsymbol{\mu}^T = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i^T$
$\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i$