Machine Learning
Lecture 10

Lecture by Dr. Shai Fine
Typeset by Steven Karas

2017-12-31
Last edited 21:13:48 2017-12-31

Disclaimer These lecture notes are based on the lecture for the course Machine Learning, taught by Dr. Shai Fine at IDC Herzliyah in the fall semester of 2017/2018. Sections may be based on the lecture slides written by Dr. Shai Fine.

Homework 4

Agenda

- Bayesian Learning & Decision Theory

1 Bayesian Decision Theory

Classifying Oranges Given a model which classifies images of oranges to either ripe or rotten.

This means we are given a priori the probabilities $\Pr(C = \text{ripe})$ and $\Pr(C = \text{rotten})$, and the class conditioned probability $\Pr(X \mid C)$ - the probability of a sensor measurement given a type of orange.

Under i.i.d. samples, we apply Bayes rule:

$$\overbrace{\Pr(C \mid x)}^{\text{posterior}} = \frac{\overbrace{\Pr(C)}^{\text{prior}}\overbrace{\Pr(x \mid C)}^{\text{likelihood}}}{\underbrace{\Pr(x)}_{\text{evidence}}}$$

1.1 Bayes Rule

Let C be a random variable with possible values $\Omega = \{\omega_1, \ldots, \omega_k\}$. Denote the *class prior* as $\Pr(C = \omega_i) = \Pr(\omega_i)$

Let $x = (x_1, \ldots, x_d)$ be the d-dimensional feature vector in feature space S, where S can be either discrete or continuous. Denote the *Class conditional probability* as $\Pr(x \mid \omega_i)$.

Denote the *evidence* as $\Pr(x) = \sum_i \Pr(x \mid \omega_i)\Pr(\omega_i)$

Denote the *posterior probability* as $\Pr(\omega_i \mid x)$. This is due to Bayes Theorem:

$$\overbrace{\Pr(\omega_i \mid x)}^{\text{posterior}} = \frac{\overbrace{\Pr(x \mid \omega_i)}^{\text{likelihood}}\overbrace{\Pr(\omega_i)}^{\text{prior}}}{\underbrace{\Pr(x)}_{\text{evidence}}}$$

Let $\alpha(x) : S \to A$ be a mapping from the feature space to some set of decisions $A = \{\alpha_1, \ldots, \alpha_a\}$. We call $\alpha(x)$ a decision rule. If $A = \Omega$, then we refer to $\alpha(x)$ as a classifier.

1.2 The MAP Rule

The Maximum APosterior Rule states that:

$$g^*(x) = \max_{\omega_i} \Pr(\omega_i \mid x) = \max_{\omega_i} \Pr(x \mid \omega_i)\Pr(\omega_i)$$

Note that because we take the maximum, they all share the same evidence, and so it's already scaled by that.

Degenerate cases If for some x, $\forall i : \Pr(x \mid \omega_i) = \Pr(x)$, then this feature vector gives us no information about the state. Also, it indicates that the decision will be based only on the prior possibilities $g^*(x) = \max_{\omega_i} \Pr(\omega_i)$.

If $\forall i \Pr(\omega_i) = 1/|\Omega|$, then the decision will be only based on the likelihood: $g^*(x) = \max_{\omega_i} \Pr(x \mid \omega_i)$.

The probability of misclassification is minimized by selecting the class having the largest posterior probability.

Proof of optimality It is possible to quantify the error given some model.

$$\begin{aligned}
\Pr_{g^*}(\text{error} \mid x) &= \Pr(g^*(x) \neq \omega \mid x) \\
&= \Pr(g^*(x) = \omega_1, \omega_2 \mid x) + \Pr(g^*(x) = \omega_2, \omega_1 \mid x) \\
&= \overbrace{\Pr(g^*(x) = \omega_1 \mid \omega_2, x) \Pr(\omega_2 \mid x)}^{p_1} \\
&\quad + \underbrace{\Pr(g^*(x) = \omega_2 \mid \omega_1, x) \Pr(\omega_1 \mid x)}_{p_2}
\end{aligned}$$

If $\Pr(\omega_1 \mid x) > \Pr(\omega_2 \mid x)$, then $g^*(x) = \omega_1$ and $p_1 = 1$, $p_2 = 0$. Therefore, $\Pr_{g^*}(\text{error} \mid x) = \Pr(\omega_2 \mid x)$.

If $\Pr(\omega_2 \mid x) > \Pr(\omega_1 \mid x)$, then $g^*(x) = \omega_2$ and $p_1 = 0$, $p_2 = 1$. Therefore, $\Pr_{g^*}(\text{error} \mid x) = \Pr(\omega_1 \mid x)$.

1.3 Generic loss function

Given some classes $\omega_1, \ldots, \omega_k$ and suppose we can make predictions $\alpha_1, \ldots, \alpha_k$.

A loss function $\lambda(\alpha_i \mid \omega_j)$ describes the loss associated with making a prediction α_i when the class is ω_j.

Risk With the estimates of $\Pr(\omega_j \mid x)$, we can compute the conditional risk associated with prediction α_i:

$$R(\alpha_i \mid x) = \sum_j \lambda(\alpha_i \mid \omega_j) \Pr(\omega_j \mid x)$$

The overall risk is the expectation of the loss:

$$R = \int R(\alpha(x) \mid x) \Pr(x)\, dx$$

Minimizing error rate classification The zero-one loss function is defined as:

$$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j) = \begin{cases} 1 & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

This gives us a conditional risk that is equivalent to the classification error:

$$\begin{aligned}
R(\alpha_i \mid x) &= \sum_{j=1}^{n} \lambda(\alpha_i \mid \omega_j) \Pr(\omega_j \mid x) \\
&= \sum_{j \neq i} \Pr(\omega_j \mid x) \\
&= 1 - \Pr(\omega_i \mid x) \\
&= \Pr(\text{error} \mid x)
\end{aligned}$$

As such, Bayes rule of $g^*(x) = \omega_i$ if $\Pr(\omega_i \mid x) > \Pr(\omega_j \mid x)$ for all $i \neq j$ gives us a minimum error rate:

$$R^* = \min_{\alpha(x)} \int R(\alpha(x) \mid x) \Pr(x)\, dx$$

Other loss functions For classification problems, 0-1 is typical, but for prediction or regression problems, other loss functions can be very useful to introduce the concept of distance between predictions.

- 0-1 loss: $\lambda_{\text{0-1}} = I[x \neq y]$
- Absolute loss: $\lambda_{\text{abs}} = |x - y|$
- Squared loss[1]: $\lambda_{\text{sq}} = (x - y)^2$

Cost based classification Let $\alpha_i = \omega_i, i = \{1, 2\}$ and $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$:

$$R(\alpha_1 \mid x) = \lambda_{11} \Pr(\omega_1 \mid x) + \lambda_{12} \Pr(\omega_2 \mid x)$$
$$R(\alpha_2 \mid x) = \lambda_{21} \Pr(\omega_1 \mid x) + \lambda_{22} \Pr(\omega_2 \mid x)$$

Under Bayes rule of $g^*(x) = \omega_1$ if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$. In other words, $(\lambda_{21} - \lambda_{11})P(\omega_1 \mid x) > (\lambda_{12} - \lambda_{22}) \Pr(\omega_2 \mid x)$.

Assuming that $\lambda_{21} > \lambda_{11}$ and $\lambda_{12} > \lambda_{22}$ (errors cost more than correct decisions), then:

$$\frac{\Pr(\omega_1 \mid x)}{\Pr(\omega_2 \mid x)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})}$$

We can restate this using Bayes rule, such that $g^*(x) = \omega_1$ if:

$$\frac{\Pr(x \mid \omega_1)}{\Pr(x \mid \omega_2)} > \frac{(\lambda_{12} - \lambda_{22}) \Pr(\omega_2)}{(\lambda_{21} - \lambda_{11}) \Pr(\omega_1)} = \theta_\lambda$$

Discriminant Functions Details on slide 18.

2 Empirical Risk Minimization

When we don't have the true conditional distributions $P(\omega_j \mid x)$, we estimate the risk of a given hypothesis function h using a training set:

$$R_{\text{emp}} = \frac{1}{n} \sum_{i=1}^{n} \lambda\left((x_i, y_i), h\right)$$

The learning paradigm utilizes the *empirical risk minimization* principle:

$$\min_{h \in H} R_{\text{emp}}(f) = \min_{h \in H} \left[\frac{1}{n} \sum_{i=1}^{n} \lambda\left((x_i, y_i), h\right) \right]$$

To prevent overfitting, complex hypotheses are penalized when computing their empirical risk using some regularization constant c and complexity function $\rho(h)$:

$$\min_{h \in H} R_{\text{emp}}(f) = \min_{h \in H} \left[\frac{1}{n} \sum_{i=1}^{n} \lambda\left((x_i, y_i), h\right) + c \cdot \rho(h) \right]$$

Discriminative vs Generative Models Discriminative models model the decision boundary $\Pr(y \mid x)$ directly, and focuses on the tail.

Typically has good performance, especially when the underlying model is misspecified. However, it is sensitive to noise, and has limited scope.

Generative models model $\Pr(x \mid y)$ and $\Pr(y)$, which gives a full density model of $\Pr(x, y)$, and focuses on the center of mass. They infer $\Pr(y \mid x)$ via Bayes rule.

This makes multiclass easy, and is robust to misclassification and noise. It can identify anomalous instances, and can return a confidence.

However, Bayes rule assumes the density models are perfect, which they never are. It also solves classification, but only through calculation of the decision boundary, rather than modeling it.

[1] This is a very useful loss function because it is derivable at all points

3 Bayesian learning models

Assume that the target concept is in hypothesis class H and it is chosen according to some prior distribution $\Pr(h)$ over $h \in H$, where the distribution is known to the learner.

Given a training set D, for each hypothesis $h \in H$, calculate the posterior probability:

$$\Pr(h \mid D) = \frac{\Pr(D \mid h) \Pr(h)}{\Pr(D)}$$

Output the hypothesis h_{MAP} with the highest probability:

$$h_{\text{MAP}} = \arg\max_{h \in H} \Pr(h \mid D)$$

3.1 Optimal Classifier

Basically, rather than taking the maximum hypothesis, weigh them according to their posterior probabilities.

$$c^*(x) = \arg\max_c \sum_{h \in H} \Pr(h \mid D) \Pr(c \mid h, x)$$

An example can be found on slide 24.

The issue with this approach is that if the hypothesis space is large (or infinite), it is not feasible to compute this.

3.2 Gibbs classifier

Choose one hypothesis $h \in H$ at random, according to the posterior probability distribution $\Pr(h \mid D)$. Use this to classify a new instance x.

The expected error of the Gibbs algorithm is at worst twice the expected error of the Bayes optimal classifier:

$$E[\text{err}_{\text{Gibbs}}] \leq 2 \cdot E[\text{err}_{\text{Bayes Optimal}}]$$

We can improve this further by sampling multiple hypotheses from $\Pr(h \mid D)$ and average their classification results.

3.3 Bagging Classifiers

Bootstrap aggregating classifiers avoid sampling from $\Pr(h \mid D)$ because it's difficult.

Given a dataset D containing m training examples, we create multiple copies D^i by drawing m samples at random with replacement from D.

The Bagging algorithm creates k bootstrap samples $D^1, \ldots, D^k$. It then trains distinct classifiers h_i on each D^i. Finally, it classifies a new instance x by classifier vote with equal weights:

$$c^*(x) = \arg\max_c \sum_{i=1}^{k} \Pr(c \mid h_i, x)$$

Note that Bagging is basically equivalent to Bayesian Average:

$$\sum_i \underbrace{\Pr(c \mid h_i, x)}_{\text{Bagging}} \approx \sum_i \underbrace{\Pr(c \mid h_i, x)}_{\text{Bayesian average}}$$

4 Next week

- SVM
- Ensemble methods
- Neural Networks & Deep Learning