# گزارش تکلیف پنجم درس الگوریتم های علوم داده


**نام و نام خانوادگی: فاطمه ترودی**

**شماره دانشجویی: ۴۰۳۴۲۲۰۴۸**

**نام استاد: دکتر سعیدرضا خردپیشه**


**نیمسال دوم ۰۴-۱۴۰۳**

# فهرست مطالب

## سوال یک. اثربخشی مهندسی ویژگی

مهندسی ویژگی یک فرآیند خلاقانه است، زیرا مستلزم **خلق** ویژگی‌های جدید از داده‌های خام است تا الگوهای پنهان را آشکار کند. برای انجام مؤثر آن، تیم‌ها باید:

- **دانش حوزه (Domain Knowledge) را به کار بگیرند:** همکاری با متخصصان برای کشف ویژگی‌های مرتبط با کسب‌وکار.
- **رویکرد مبتنی بر فرضیه داشته باشند:** به جای ایجاد کورکورانه ویژگی‌ها، فرضیه‌هایی را در مورد آن‌ها مطرح کنند.
- **از یک فرآیند چرخه‌ای پیروی کنند:** از تحلیل اکتشافی داده‌ها (EDA) شروع کنند، ویژگی‌ها را بسازند، آن‌ها را ارزیابی و سپس این چرخه را تکرار کنند.

---

## سوال دو. روش‌های کدگذاری متغیرهای دسته‌بندی

**(الف) انواع روش‌ها و مقایسه آن‌ها:**

| روش | توضیح | مزایا | معایب |
|---|---|---|---|
| **One-hot Encoding** | برای هر دسته یک ستون باینری می‌سازد. | ساده و قابل فهم. | با تعداد زیاد دسته‌ها، منجر به ابعاد بسیار زیاد و پراکنده می‌شود. |
| **Label Encoding** | هر دسته را به یک عدد صحیح نگاشت می‌کند. | کارآمد از نظر حافظه. | یک رابطه ترتیبی (ordinal) غلط ایجاد می‌کند که می‌تواند مدل را گمراه کند. |
| **Target Encoding** | هر دسته را با میانگین متغیر هدف جایگزین می‌کند. | بسیار قدرتمند و پیش‌بینی‌کننده. | به شدت مستعد **بیش‌برازش (Overfitting)** است. |
| **Hash Encoding** | از یک تابع هش برای نگاشت دسته‌ها به تعداد ثابتی از ستون‌ها استفاده می‌کند. | کارآمد از نظر حافظه و حل مشکل ابعاد زیاد. | ممکن است به دلیل **برخورد هش (hash collision)** اطلاعات از بین برود. |

(ب) محدودیت One-hot Encoding:

اصلی‌ترین محدودیت، افزایش شدید ابعاد (curse of dimensionality) است که در نتیجه، مدل کند و از نظر حافظه پرهزینه می‌شود.

(ج) مزیت Label Encoding بر One-hot:

زمانی که داده‌ها رابطه ترتیبی واقعی دارند (مانند اندازه‌های لباس S, M, L) و هنگام استفاده از مدل‌های مبتنی بر درخت (مانند Random Forest) که به ترتیب عددی حساس نیستند، Label Encoding کارآمدتر است.

(د) جلوگیری از بیش‌برازش در Target Encoding:

با استفاده از اعتبارسنجی متقابل (Cross-validation)، داده‌های آموزشی را تقسیم کرده و میانگین هدف را تنها بر روی داده‌های خارج از بخش فعلی محاسبه می‌کنیم. این کار از حفظ کردن (memorizing) اطلاعات هدف جلوگیری می‌کند.

(ه) مزایا و معایب Hash Encoding:

این روش با نگاشت تعداد زیادی دسته به تعداد کمتری از ستون‌ها، مشکل پراکندگی را حل می‌کند. اما مشکل اصلی آن برخورد هش است که باعث می‌شود دو دسته متفاوت به یک ستون یکسان نگاشت شوند.

---

**سوال سه**

Category Embeddingها هر دسته را به یک **بردار متراکم و کوچک** تبدیل می‌کنند که در طول آموزش یک شبکه عصبی یاد گرفته می‌شود. این روش بهتر از روش‌های سنتی عمل می‌کند، زیرا:

- **روابط ذاتی را یاد می‌گیرد:** دسته‌های مشابه در فضای برداری به هم نزدیک می‌شوند.
- **تعداد زیاد دسته‌ها را به خوبی مدیریت می‌کند:** از افزایش ابعاد و پراکندگی جلوگیری می‌کند.
- **نمایش بهینه را به طور خودکار کشف می‌کند.**

---

**سوال چهار. گرد کردن ویژگی‌های عددی**

**مزیت:** می‌تواند **نویز را کاهش داده** و به مدل کمک کند تا بر روی الگوهای اصلی تمرکز کند.

**ریسک:** می‌تواند باعث **از دست رفتن اطلاعات مهم** شود، به خصوص اگر تفاوت‌های کوچک بین مقادیر، معنای مهمی داشته باشند.

---

**سوال پنج. غیرخطی بودن و ویژگی‌های چرخه‌ای**

**(الف) کمک به مدل‌های خطی:**

- **ویژگی‌های چندجمله‌ای:** با ایجاد ویژگی‌هایی مانند $X^2$ و $X^3$ به مدل اجازه می‌دهند تا منحنی‌ها را برازش دهد.
- **ویژگی‌های تعاملی:** با ضرب ویژگی‌ها در هم (مثلاً $X1 \times X2$)، مدل می‌تواند اثر متقابل آن‌ها را یاد بگیرد.

(ب) ویژگی‌های چرخه‌ای (Cyclical Features):

برای داده‌های زمانی مانند ساعت روز، به جای یک عدد، از توابع سینوس و کسینوس استفاده می‌کنیم تا به مدل نشان دهیم که پایان و ابتدای یک چرخه به هم نزدیک هستند.

---

**سوال شش**

**TF-IDF** برای **برجسته کردن کلمات مهم** در یک سند استفاده می‌شود و به مدل کمک می‌کند تا به جای کلمات رایج، روی کلمات کلیدی متمرکز شود.

**کاهش ابعاد** (مانند PCA) برای داده‌های متنی پرکاربرد است، زیرا:

- **ابعاد زیاد را کاهش می‌دهد:** باعث سرعت بخشیدن به آموزش و کاهش مصرف حافظه می‌شود.
- **نویز را حذف می‌کند:** با تمرکز بر مهم‌ترین اطلاعات.

---

## سوال هفت. شبکه‌های عصبی و مهندسی ویژگی

شبکه‌های عصبی به مهندسی ویژگی دستی کمتری نیاز دارند، زیرا:

- **یادگیری سلسله‌مراتبی ویژگی:** لایه‌های اولیه ویژگی‌های ساده را یاد می‌گیرند و لایه‌های بعدی آن‌ها را ترکیب می‌کنند تا ویژگی‌های پیچیده‌تر را بسازند.
- **قابلیت غیرخطی بودن:** خود شبکه به صورت خودکار قادر به یادگیری روابط غیرخطی پیچیده است.
- **یادگیری خودکار Embeddingها:** می‌تواند بهترین نمایش برای ویژگی‌های دسته‌بندی را به طور خودکار پیدا کند.